SHRINKAGE IN QUICKEST CHANGE DETECTION, MULTICHANNEL PROFILE MONITORING, AND UNCERTAINTY QUANTIFICATION

A Thesis
Presented to
The Academic Faculty

by

Yuan Wang

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
H. Milton Stewart School of Industrial and Systems Engineering

Georgia Institute of Technology
August 2016

SHRINKAGE IN QUICKEST CHANGE DETECTION, MULTICHANNEL PROFILE MONITORING, AND UNCERTAINTY QUANTIFICATION

Approved by:

Dr. Yajun Mei, Advisor
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Jeff Wu, Advisor
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Kamran Paynabar
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Yao Xie
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Ray-bing Chen
Department of Statistics
National Cheng Kung University

Date Approved: April 27, 2016

To my beloved parents,

for their unconditional love and support.

ACKNOWLEDGEMENTS

The past five years are the most enjoyable part of my life. I am greatly indebted to many remarkable people. Without their support and help, the completion of my dissertation would be impossible.

First and foremost, my deepest gratitude must go to my advisor, Prof. Yajun Mei, for his countless guidance and continuously support in my research and career development. I couldn't be more appreciating his incredible dedication in advising me during my doctoral studies. Whenever I need help, he always guides me through the maze with great patience.

Second, I owe great gratitude to my co-advisor, Prof. Jeff Wu, for being an incredible mentor in academics, life, and career. His enthusiasm and broad knowledge have inspired and guided me throughout the study. I really appreciate the tremendous guidance and support from Prof. Wu during my whole Ph.D. journey.

Moreover, I would like to express my deepest appreciation to my committee members, Prof. Kamran Paynabar, Prof. Ray-Bing Chen, and Prof. Yao Xie, for their generous help, efforts, invaluable discussions and suggestions on my doctoral study and dissertation. Furthermore, their tremendous dedication deeply affects me, which nurtures me to become a better statistician.

I will always remember those excellent and insightful courses I took during the past five years. I would like to thank Prof. Roshan Joseph Vengazhiyil, Prof. Xiaoming Huo, Prof. Jianjun Shi, Prof. Sigrun Andradottir and Prof. Enlu Zhou for teaching me great courses. Their help provides me a wide spectrum of knowledge and experience on statistics and industrial engineering.

I would also like to extend a special thanks to Prof. Hongquan Xu at UCLA. Prof. Xu has advised my undergraduate research during the summer of 2010, and he is the first one who introduced me to statistical research and he recommended me to Georgia Tech.

I am very thankful for my friends, officemates, alumni, and classmates: Niao He, Jia

Yan, Yun Liu, Li Gu, Heng Su, Dianpeng Wang, Simon Mak, Chih-Li Sung, David Zhao, Fang Cao, Chengliang Zhang, Seonghye Jeon, Huizhu Wang, Evren Gul, Chitta Ranjan, Weijun Ding, Qianyi Wang, Qiushi Chen, Xinyu Min, Huizhi Xie, Huan Yan, Zhi Han, Ralph Yuan, Fan Ye, Yibiao lv, Yi Xiao, Jie Chen and many others I could not name them all. It is them who make me feel that our School of ISyE is like a warm big family. They help me a lot on both life and career, and provide me endless source of inspiration and happiness in my student life.

Finally but definitely not least, I would like to thank my beloved parents, who brought me to this wonderful world, and provide unlimited love and support through my whole life. I am extremely blessed to have my special one, Suo Yang, who accompanies me throughout my Ph.D. journey, and the unknown but exciting adventures ahead of me. This dissertation is dedicated to them.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

In the information age, many real-world applications such as biosurveillance, manufacturing systems, physical and computer experiments often involve data that are massive, high-dimension or have complicated structures. In some cases it is cheap to collect large-scale data, while in other cases it may be costly or time-consuming to collect them. In either case, it is often non-trivial to extract information from these types of data to make useful decisions.

This dissertation makes methodology contributions to three important subfields of statistics: (i) Large-scale multi-stream quickest change detection, (ii) multichannel profile monitoring and (iii) global optimization of expensive functions. A common feature of the thesis work is the use of shrinkage to the respective subfields to address the challenges of high-dimensional or complicated data. However, since different subfields and applications have different features and challenges, details of the shrinkage techniques vary with the subfield.

This dissertation consists of three chapters. In Chapter 1, we study the problem of online monitoring large-scale data streams, which has many important applications from biosurveillance and quality control to finance and security in modern information age. While many classical quickest change detection methods can be extended from one-dimensional to any K-dimensional, their performances are rather poor when monitoring large K of data streams. This motives us to investigate the effects of dimensionality on the performance of quickest change detection methods. We found out through theoretical analysis that the classical quickest change detection methods often over-emphasize the first-order term of the detection delays and overlook the second-order terms of the detection delays, where the latter often increases linearly as a function of the dimension K. When K is large (e.g., hundreds), the second-order term of the detection delay will likely be comparable to the first-order term, which implies that the nice first-order asymptotic optimality properties

have little practical meaning for large K. We propose a novel approach to lessen the dimensionality effects by introducing some shrinkage estimators of the unknown post-change parameters. In addition, we also illustrate the challenge of Monte Carlo simulation of the average run length to false alarm in the context of online monitoring large-scale data streams. Most of the material in Chapter 1 was published in 2015 in IEEE Transactions on Information Theory.

In Chapter 2, we consider the problem of monitoring multichannel profiles that has important applications in manufacturing systems improvement. A concrete motivating example of this work is from a forging process, in which multichannel load profiles measure exerted forces in each column of the forging machine. While various methods have been developed for univariate profile monitoring, they often cannot easily be extended to multichannel profiles. There are two main challenges when monitoring multichannel profiles. The first one is that profiles are high-dimensional functional data with intrinsic inner- and inter-channel correlations, and the second, probably more fundamental, challenge is that the functional structure of multi-channel profiles might change over time, and thus the dimension reduction method should be capable of taking into account the potential unknown change. We develop a novel thresholded multivariate principal component analysis (PCA) method for multi-channel profile monitoring. Our proposed method consists of two steps of dimension reduction: It first applies the functional PCA to extract a reasonable large number of features under the in-control state, and then uses the shrinkage techniques to functional PCAs to further select significant features capturing profile information in the out-of-control state. The choice of tuning parameter for soft-thresholding is provided based on asymptotic analysis, and extensive simulation studies are conducted to illustrate the efficacy of our proposed methodology.

In Chapter 3, we study the problem of global optimization of expensive functions. In modern physical and computer experiments, one often deals with expensive functions in the sense that it may take days or months to evaluate their values at a single input setting. An important problem is how to choose an appropriate setting of the input variables so as to optimize the output. To tackle this question, our proposed method involves two main

components: one is the construction of a surrogate model to approximate the true function with much cheaper computation, and the other is the determination of a new input setting for function evaluation based on the surrogate model. After iteratively updating these two components, we optimize the latest surrogate model, which yields the approximation to the optima of the original expensive function. To be specific, we propose an adaptive Radial Basis Function (RBF) based global optimization framework via uncertainty quantification. For the surrogate model, we construct an adaptive RBF-based normal mixture Bayesian surrogate model, where the parameters in the RBFs can be adaptively updated each time a new point is explored. It is crucial to employ the normal mixture Bayesian structure which leads to a more stable surrogate model and avoid over-fitting. Its use can be regarded as a ridge-type regression estimate of model coefficients. For the selection of input setting, we propose a novel criterion to assess the input setting based on the surrogate model, and we choose the inputs that maximize the criterion. Our criterion incorporates the expected improvement (EI) of the function prediction to effectively identify promising areas for the global optima, and its uncertainties to efficiently explore the unknown regions. We conduct numerical studies with standard test functions to understand and compare the empirical performance of our proposed method with a prominent existing method.

CHAPTER I

LARGE-SCALE MULTI-STREAM QUICKEST CHANGE DETECTION VIA SHRINKAGE POST-CHANGE ESTIMATION

1.1 Introduction

The problem of online monitoring large-scale data streams has many important applications from biosurveillance and quality control to finance and security in modern information age when the rapid development of sensing technology allows one to generate large-scale real-time streaming data. In many scenarios, one is often interested in the early detection of a "trigger" event when "sensors" are deployed to monitor the changing environments over time and space, see Lawson and Kleinman [32]. From the theoretical or methodological viewpoint, this is a quickest change detection or sequential change-point detection problem, where the case of monitoring $K = 1$ data stream has been extensively studied in the past several decades, see the books by Basseville and Nikiforov [3] and Poor and Hadjiliadis [44] for the review. Also see Page [39], Shiryaev [52], Lorden [33], Pollak [42], Moustakides [37], Lai [31] for some early classical contributions. In addition, the case of online monitoring a not so large number K (e.g., tens) of data streams has also been studied in the literature, see Lorden and Pollak [34], Tartakovsky et al. [54], Zamba and Hawkins [59], Veeravalli and Bangerjee [55].

Unfortunately research on the problem of online monitoring a large number K (e.g., hundreds or more) of data streams is rather limited, see Siegmund [53] and the discussions therein. While many classical quickest change detection methods are based on likelihood ratio statistics, and can be extended from one-dimensional to K-dimensional, their performances are rather poor when monitoring a large number K of data streams, despite holding the so-called first-order asymptotic optimality properties for any fixed dimensional K in the sense of asymptotically minimizing the detection delay for each and every possible post-change hypothesis as the average run length (ARL) to false alarm constraint goes to

∞. The main reason is that these classical quickest change detection methods often over-emphasize the first-order performance for *each and every possible* post-change hypothesis in the K-dimensional space, and thus the price they paid is on the second-order terms of the detection delays which are often linearly increasing as a function of K. This is not an issue when the number K of data streams is small, but it has a severe effect when K is large (e.g., hundreds): under a reasonable practical setting, the second-order term of the detection delay will likely be comparable to the first-order term, which implies that the nice first-order asymptotic optimality properties have little practical meaning for large K! This led Mei [35] to raise an open problem whether one can develop new methods that can reduce the coefficient in the second-order term of the detection delay from K to a smaller number to yield quicker detection.

The primary objective of this paper is to tackle this open problem, and propose a systematic approach to develop efficient methodologies for online monitoring a large number K of independent data streams. Our proposed methods do not aim for each and every possible post-change hypothesis in the K-dimensional space, and the main assumption we make is that for each individual local data stream, either there are no local changes, or there is a local change that is larger than some pre-specified lower bounds. The key novelty of our proposed methodologies is to apply shrinkage estimators to incorporate such prior knowledge of the post-change hypothesis to develop efficient quickest change detection methodologies. To illustrate our main ideas, we will focus on the problem of monitoring K independent normal data streams with possible changes in the means of some data streams, and two different scenarios will be investigated: one is the sparse post-change case when the unknown number of affected data streams is much smaller than the total number of data streams, and the other is when all local data streams are affected simultaneously although not necessarily identically, i.e., different local data streams may have different *unknown* post-change mean parameters. It is useful to think that for a given total information for changing event, the former scenario corresponds to the case of a few "large" local changes, whereas the latter scenario corresponds to the case of "relatively small" local changes in all data streams. Given the same total information of changing event, the classical quickest change

detection methods will have similar (first-order) performance under these two scenarios, although intuitively one may feel that these two scenarios should be different. Our proposed methods combine the hard thresholding estimators with the linear shrinkage estimators to simultaneously estimate unknown post-change mean parameters, and will indeed show that these two scenarios should be treated differently. In the process of investigating the properties of the proposed methods, we also demonstrate the challenge of Monte Carlo simulation of the average run length to false alarm for large dimensional K due to the curse of dimensionality, which seems to be overlooked in the quickest change detection literature.

Note that the usefulness of shrinkage or thresholding in high-dimensional data is well-known in the modern off-line statistical research since the pioneering work of James and Stein [23], also see Candés [8] and references therein. However, the application of shrinkage or thresholding to quickest change detection is rather limited. Unlike other off-line works that deal with high-dimensional statistics, the asymptotic analysis in this paper fixes the dimension K (or the number of data streams) as the ARL to false alarm is taken to infinity. Our aim is on the development of asymptotic results that are useful for the practical setting, and thus our focuses are on the effects of the dimension K on the second-order term of the detection delays, and on how shrinkage or thresholding can lessen such effects. As far as we know, it remains an open problem in quickest change detection when the dimension K is taken to infinity.

In the present paper, we will demonstrate how to combine shrinkage estimators with the classical Shiryaev-Roberts procedure to yield an efficient global monitoring scheme. Note that the Shiryaev-Roberts procedure is chosen as a demonstration here, since it allows us to simplify our mathematical arguments by borrowing the results in Lorden and Pollak [34] that develops the Shiryaev-Roberts-Robbins-Siegmund (SRRS) scheme based on the method of moments (MOM) estimators or the maximum likelihood estimators (MLE) of unknown post-change parameters. Besides the different estimators of unknown post-change parameters, another main difference between our research and Lorden and Pollak [34] is that we explicitly investigate the effect of the number of data streams on the detection delay performance of the schemes. We want to emphasize that our use of shrinkage estimators can

easily be combined to other popular quickest change methods such as the CUSUM procedure proposed by Page [39] from the methodology or algorithm point of view, although the corresponding theoretical asymptotic analysis seems to be nontrivial. Hopefully our useful of shrinkage estimation opens new directions to develop more efficient methodologies for online monitoring of large-scale or high-dimensional data streams.

From the information theory viewpoint, the asymptotic performance of our proposed shrinkage-based schemes is characterized by the new information number defined in (1.4.4) below. In a simple setting for normal distributions when the ω_k's are the smallest meaningful bounds on the post-change mean parameters μ_k's, the new information number has the form of $\frac{1}{2}\sum_{k:|\mu_k|>\omega_k}(\mu_k)^2$, whereas the classical Kullback-Leibler divergence is $\frac{1}{2}\sum_{k=1}^{K}(\mu_k)^2$. Thus our proposed new information number can be thought of as the shrinkage approximation of the classical Kullback-Leibler divergence between pre-change and post-change distributions. In the context of monitoring large-scale data streams, we feel that our proposed new information number in (1.4.4) provides more meaningful bounds than the classical Kullback-Leibler divergence, since it takes into account of the second-order term of the detection delay performance and the spatial uncertainty associated with which local data streams are affected.

We should acknowledge that Xie and Siegmund [57] studies a similar problem by taking a semi-Bayesian approach under the assumption that the fraction of affected data streams is known. Here we did not make such an assumption, and our formulation assumes that the lower bound of the post-change parameters are given, i.e., we are only interested in detecting certain large local changes for individual local data streams. In addition, Tartakovsky et al. [54] and Mei [35] consider the special case when all post-change parameters for affected data streams were identical or completely specified. Here our underlying assumption is that the post-change parameters are unknown and not necessarily identical. In addition, the problem of monitoring $K > 1$ data streams is also studied in the offline setting when the full information is available during decision-making, e.g. Zhang, Siegmund, Ji and Li [61], and Cho and Fryzlewicz [12]. Our setting here is online where we observe the data sequentially over time, and we cannot use future observations to make current decision.

The remainder of this paper is organized as follows. In Section 1.2, we state the mathematical formulation of monitoring $K > 1$ data streams and review shrinkage estimators in offline point estimation that will be used later. In Section 1.3, we propose our shrinkage-based monitoring scheme for the problem of online monitoring of independent normal data streams with possible changes in some of the means. Section 1.4 develops asymptotic properties of our proposed monitoring schemes. In Section 1.5, we report numerical simulation results to illustrate the usefulness of our proposed shrinkage-based schemes and the challenge of Monte Carlo simulation of the average run length to false alarm in the context of online monitoring large-scale data streams. Section 1.6 contains some concluding remarks.

1.2 Problem Formulation and Background

1.2.1 Problem Formulation

Assume we are monitoring K independent normal data streams in a system. Denote by $X_{k,n}$ the observation of the k-th data stream at time n for $k = 1, ..., K$ and $n = 1, 2, ...$ The $X_{k,n}$'s are assumed to be independent not only over time within each data stream, but also among different data streams. Initially, all $X_{k,n}$'s are independent and identically distributed (iid) $N(\mu_0, 1)$ random variables. At some unknown time $\nu \in \{1, 2, 3, \ldots\}$, an event may occur to the system, and affect some data streams in the sense that the distribution of the $X_{k,n}$'s may change to $N(\mu_k, 1)$ for $n = \nu, \nu+1, ...,$ if the k-th data stream is affected for $k = 1, \ldots, K$. To simplify our notation, here the post-change mean $\mu_k = \mu_0$ implies that the corresponding data stream is not affected, whereas $\mu_k \neq \mu_0$ corresponds to an affected data stream. Following the literature of quickest change detection, we assume that the pre-change mean μ_0 is completely specified, and without loss of generality, we assume $\mu_0 = 0$, as otherwise we can monitor $X_{k,n} - \mu_0$ instead of $X_{k,n}$'s themselves. Thus μ_0 and 0 are interchangeable below for normal distributions.

In this article, we tackle the case when the post-change means μ_k's are only partially specified, e.g., we do not know which data streams are affected and do not know the exact values of the post-change means μ_k's for affected data streams. In practical situations when monitoring large-scale data streams, one is often interested in only detecting "big" local

changes in individual data streams. This motivates us to assume that the post-change hypothesis set for the post-change mean vector $\mu = (\mu_1, \ldots, \mu_K)^T$ is given by

$$\Omega = \{\mu \neq \mathbf{0} \in \mathcal{R}^K : \sum_{k=1}^{K} |\mu_k| 1\{|\mu_k| \leq \omega_k\} = 0\}, \tag{1.2.1}$$

where the lower bounds ω_k's are pre-specified positive constants that are the smallest difference meaningful for detection. The post-change hypothesis set Ω in (1.2.1) implies that for any local data stream, either there are no local changes (i.e., $\mu_k = 0$), or there is a big local change (i.e., $|\mu_k| > \omega_k$). In addition, $\mu \neq \mathbf{0}$ implies that at least one $\mu_k \neq 0$, i.e., at least one data stream should be affected under the post-change hypothesis. Also note that the post-change hypothesis set Ω in (1.2.1) assumes the true post-change mean $\mu_k \neq \pm\omega_k$ for any k. This is a technical assumption to simplify our theoretical analysis, since otherwise careful arguments are needed to take care of those data streams with $|\mu_k| = \omega_k > 0$ which could be thought of as affected data streams only with probability $1/2$. For any given post-change mean vector μ, it is natural to define the number of affected data streams as $r = \sum_{k=1}^{K} 1\{\mu_k \neq 0\}$ where $1\{A\}$ is the indicator function of event A. Clearly, when $\mu \in \Omega$ in (1.2.1), this becomes

$$r = \sum_{k=1}^{K} 1\{|\mu_k| > \omega_k\}, \tag{1.2.2}$$

which will play an important role on the detection delay performance of quickest change detection schemes in our context. Note that the main scheme in Xie and Siegmund [57] assumes that the number of affected data streams, or the r value in (1.2.2), is known and the lower bound $\omega_k = 0$ for all k. In this article, we assume that the lower bounds ω_k's in (1.2.1) are known **positive** constants for all $k = 1, \ldots, K$. Two scenarios will be studied: one is the sparse post-change hypothesis case when the value r in (1.2.2) is an unknown constant that is much smaller than K, and the other is when $r = K$, i.e., when all data streams are affected simultaneously.

To provide a more rigorous mathematical formulation, denote by $\mathbf{P}_{\mu,\nu}$ and $\mathbf{E}_{\mu,\nu}$ the probability measure and expectation of $\{(X_{k,1},\ X_{k,2},...)\}_{k=1}^{p}$ when the change occurs at time ν and the true post-change mean vector $\mu = (\mu_1, \ldots, \mu_p)^T$. Denote by $\mathbf{P}_\infty$ and $\mathbf{E}_\infty$

the same when no change occurs, i.e., the change-time $\nu = \infty$. Loosely speaking, we want to develop an online global monitoring scheme that can raise a true alarm as soon as possible when the event occurs while controlling the global false alarm rate. Mathematically, an online global monitoring scheme is defined as a stopping time T, which is an integer-valued random variable. The event $\{T = n\}$ represents that we will raise an alarm at time n at the global level and declare that a change occurs somewhere in the first n time steps. Note that the decision $\{T = n\}$ is only based on the observations $X_{k,i}$'s up to time n.

The standard minimax formulation of quickest change detection problem can then be formally stated as follows: Find a stopping time T that asymptotically minimizes the "worst-case" detection delay proposed in Lorden [33]

$$D_{\mu}(T) = \sup_{1 \leq \nu < \infty} \text{ess sup } \mathbf{E}_{\mu,\nu}(T - \nu + 1 | T \geq \nu, \mathcal{F}_{\nu-1})$$

for all possible post-change mean vectors $\mu \in \Omega$ in (1.2.1) subject to the constraint on the average run length (ARL) to false alarm

$$\mathbf{E}_{\infty}(T) \geq A. \tag{1.2.3}$$

Here $\mathcal{F}_{\nu-1}$ denotes all information up to time $\nu - 1$, and the constraint $A > 0$ in (1.2.3) is pre-specified.

1.2.2 Review of Shrinkage Estimation

Let us now review some well-known fact regarding offline shrinkage estimation procedures, which will be used in our proposed methodologies for online monitoring $K > 1$ data streams in the next section. Suppose that there are $K \geq 3$ independent normal random variables, say, $\{Y_1, \ldots, Y_K\}$, where $Y_k \sim N(\mu_k, \sigma^2)$ with unknown mean μ_k and known variance σ^2 for $k = 1, \ldots, K$. Suppose we are interested in estimating the K-dimensional mean vector $\mu = (\mu_1, \ldots, \mu_K)^T$ and want to find a good estimator $\widehat{\mu} = (\widehat{\mu}_1, \cdots, \widehat{\mu}_K)^T$ under the mean squared error (MSE) criterion $MSE(\widehat{\mu}) = \mathbf{E}||\widehat{\mu} - \mu||^2 = \mathbf{E}\left(\sum_{k=1}^{K}(\widehat{\mu}_k - \mu_k)^2\right)$.

It is trivial to see that the method of moment estimator (MOM) or maximum likelihood estimator (MLE) of μ_k is $\widehat{\mu}_k^{MLE} = Y_k$ for $k = 1, \ldots, K$, since each μ_k corresponds to only one normal variable Y_k. A surprising result in a remarkable paper by James and Stein [23]

is that there are uniformly better estimators than MOM or MLE in the sense of smaller MSE when simultaneously estimating $K \geq 3$ unknown parameters! Since then shrinkage estimation has become a basic tool in the analysis of high-dimensional data, especially when the object to estimate holds sparsity properties.

Many kinds of shrinkage estimators have been developed in the literature, see Candés [8] for the review and more references. Below we will review two kinds of shrinkage estimators that will be used in our proposed quickest change detection schemes. The first one is the linear shrinkage estimator of μ_k's defined by

$$\begin{aligned} \widehat{\mu}_k &= aY_k + (1-a)\zeta \\ &= aY_k + b, \end{aligned} \tag{1.2.4}$$

where $0 \leq a \leq 1$ is the shrinkage factor, ζ is a pre-specified real-valued constant (e.g., $\zeta = 0$), and $b = (1-a)\zeta$. This corresponds to shrinking the observed vector $(Y_1, \cdots, Y_K)^T$ to the pre-specified vector $(\zeta, \cdots, \zeta)^T$ as the shrinkage factor a goes to 0 (note that in a more general setting, ζ can be different for different k). Observe that the linear shrinkage estimator $\widehat{\mu}_k$ in (1.2.4) has the common shrinkage factor a for all k, and intuitively this works well when all true μ_k's are nonzero, or better yet, have similar values. The second kind of shrinkage estimator is the hard-thresholding estimator defined by

$$\widehat{\mu}_k = \begin{cases} Y_k & \text{if } |Y_k| \geq \omega_k \\ \mu_0 = 0 & \text{if } |Y_k| < \omega_k \end{cases}. \tag{1.2.5}$$

Intuitively, the hard-thresholding estimator in (1.2.5) works when only a (small) subset of μ_k's are different from $\mu_0 = 0$. In such scenario, it makes more sense to shrinking non-significant MOM or MLE estimators of μ_k's directly to 0. Indeed, the optimality properties of hard-thresholding estimators in (1.2.5) were established in the context of offline point estimation, see, for example, Donoho and Johnstone [14].

1.3 Our Proposed Monitoring Schemes

In the problem of online monitoring of K independent normally distributed data streams with possible mean changes, if we completely knew each and every post-change parameter

μ_k, then many classical quickest change detection procedures for monitoring one-dimensional data stream can be easily adapted to develop global monitoring schemes, and one of them is the well-known Shiryaev-Roberts procedure (Shiryaev [52] and Roberts [49]) that can be defined as follows in our context. Let $\Lambda_{n,m}^{SR}$ be the likelihood ratio statistic of all observations up to time n in the problem of testing H_0 : *no change* against H_1 : *a change occurs at time* $m(\leq n)$, i.e.,

$$\Lambda_{n,m}^{SR} = \prod_{\ell=m}^{n} \prod_{k=1}^{K} \frac{f_{\mu_k}(X_{k,\ell})}{f_{\mu_0}(X_{k,\ell})}, \tag{1.3.1}$$

where $f_\mu(\cdot)$ is the probability density function of $N(\mu, 1)$. At time n, the Shiryaev-Roberts procedure computes the global monitoring statistics

$$R_n^{SR} = \sum_{m=1}^{n} \Lambda_{n,m}^{SR}, \tag{1.3.2}$$

which can be thought of as assigning a uniform prior on the potential change-point values $\nu = m \in \{1, 2, \cdots, n\}$. Then the Shiryaev-Roberts procedure raises a global alarm at time

$$N_B^{SR} = \inf\{n \geq 1 : R_n^{SR} \geq B\}, \tag{1.3.3}$$

where the threshold $B > 0$ is chosen to satisfy the ARL to false alarm constraint in (1.2.3).

When the post-change parameters μ_k's are unknown, one natural possibility is to replace them by their corresponding estimators from the observed data. In the quickest change detection literature, it is standard to use MLE or MOM to estimate the unknown post-change parameters, though there are generally two different approaches, depending on whether or not to use the same estimate for all $n - m + 1$ post-change parameters μ_k's for $\ell = m, m + 1, \ldots, n$ in the likelihood ratio $\Lambda_{n,m}^{SR}$ in (1.3.1) at time $n(\geq m)$. The first one is to replace all $n - m + 1$ μ_k's by the same estimator based on all observations from the putative change-point time m to the current time step n, and thus it often leads to the generalized likelihood ratio type statistic, see Xie and Siegmund [57].

The second approach, adopted by Lorden and Pollak [34], is to use different estimates to the $n - m + 1$ μ_k's. To be more concrete, for each $k = 1, \ldots, K$, Lorden and Pollak [34]

considers $n-m+1$ MLE/MOM estimates of μ_k :

$$\begin{aligned} \widehat{\mu}_{k,m,\ell} &= \bar{X}_{k,m,\ell} \\ &= \begin{cases} \frac{X_{k,m}+\ldots+X_{k,\ell-1}}{\ell-m}, & \text{if } \ell = m+1,\ldots,n \\ \mu_0 = 0, & \text{if } \ell = m \end{cases} \end{aligned} \tag{1.3.4}$$

and then proposes to plug these $\widehat{\mu}_{k,m,\ell}$'s into (1.3.1)-(1.3.3) to yield the quickest change detection scheme. It is important to note that at time ℓ, the estimate $\widehat{\mu}_{k,m,\ell} = \bar{X}_{k,m,\ell}$ in (1.3.4) only uses the observations, $X_{k,m},\ldots,X_{k,\ell-1}$, to estimate μ_k at time ℓ, which allows one to reserve the observation $X_{k,\ell}$ only for detection of a change. By doing so, we keep two important properties of $\Lambda^{SR}_{n,m}$ in (1.3.1): (i) the recursive form $\Lambda^{SR}_{n,m} = \Lambda^{SR}_{n-1,m}\prod_{k=1}^{K}[f_{\mu_k}(X_{k,n})/f_{\mu_0}(X_{k,n})]$, and (ii) the nice property of $\mathbf{E}_\infty(\Lambda^{SR}_{n,m}) = 1$ which leads to a useful fact that $R^{SR}_n - n$ is a martingale under the pre-change hypothesis. Lorden and Pollak [34] termed their scheme as Shiryaev-Roberts-Robbins-Siegmund (SRRS) scheme, as similar idea has been used earlier in Robbins and Siegmund [48] for sequential hypothesis testing problems. Below the scheme of Lorden and Pollak [34] will be called as the original SRRS scheme, and will be denoted by N^{orig}_B. It was shown in Lorden and Pollak [34] that the original SRRS scheme N^{orig}_B is first-order asymptotically optimal when monitoring $K = 1$ data stream as the ARL to false alarm constraint A in (1.2.3) goes to ∞. After a careful analysis, it can also be shown that the first-order asymptotic optimality properties of N^{orig}_B can be extended for any fixed dimension K, but unfortunately the second-order term of the detection delay of the original SRRS scheme N^{orig}_B is a linear function of K. In other words, the original SRRS scheme N^{orig}_B of Lorden and Pollak [34] suffers the same problem of many classical schemes mentioned in Mei [35] that the coefficient of the second-order term of detection delay is of order K, and thus its first-order asymptotic optimality properties can be meaningless in the practical setting of monitoring large-scale data streams.

In this paper, we propose to develop a global monitoring scheme by combining the shrinkage estimators with the SRRS scheme of Lorden and Pollak [34]. Our motivation is fueled by the fact that we need to estimate K post-change means μ_k's simultaneously: if we let $Y_k = \widehat{\mu}_{k,m,\ell}$ in (1.3.4) for all $k = 1,\ldots,K$, then existing research in the offline point estimation suggests that the shrinkage estimators in (1.2.4) or (1.2.5) should lead a

better estimation of the true unknown post-change means μ_k's, which might lead to a better quickest change detection scheme.

Inspired by the linear shrinkage estimator in (1.2.4) and the hard thresholding estimator in (1.2.5), we propose a systematic approach that performs the linear shrinkage for values of MLE/MOM $\bar{X}_{k,m,\ell}$'s in (1.3.4) that are not thresholded. Specifically, we propose to consider the shrinkage estimators of the form

$$\widehat{\mu}_{k,m,\ell} = \begin{cases} a\bar{X}_{k,m,\ell} + b & \text{if } \ell = m+1, \cdots, n, \text{ and} \\ & |\bar{X}_{k,m,\ell}| \geq \omega_k \\ c & \text{otherwise.} \end{cases}, \tag{1.3.5}$$

where a, b, c are three constants to be specified later. Note that $a = 1, b = 0$ and $c = 0$ correspond to the hard-thresholding estimators in (1.2.5), which will be shown later to be one of reasonable good choices under the post-change hypothesis Ω in (1.2.1).

Our proposed shrinkage-based SRRS schemes are defined by plugging the shrinkage/thresholding estimators $\widehat{\mu}_{k,m,\ell}$ in (1.3.5) into (1.3.1)-(1.3.3). To be more concrete, define

$$\begin{aligned} \Lambda_{n,m} &= \prod_{\ell=m}^{n} \prod_{k=1}^{p} \frac{f_{\widehat{\mu}_{k,m,\ell}}(X_{k,\ell})}{f_{\mu_0}(X_{k,\ell})} \\ &= \Lambda_{n-1,m} \prod_{k=1}^{p} \frac{f_{\widehat{\mu}_{k,m,n}}(X_{k,n})}{f_{\mu_0}(X_{k,n})} \text{ for } n > m, \end{aligned} \tag{1.3.6}$$

where $\Lambda_{n,n} = 1$ for all $n = 1, 2, \ldots,$ and

$$R_n = \sum_{m=1}^{n} \Lambda_{n,m}, \tag{1.3.7}$$

with $R_1 = 1$. Then our proposed shrinkage-based SRRS scheme raises an alarm at the first time

$$N_B = \inf\{n \geq 1 : R_n \geq B\}, \tag{1.3.8}$$

where $B > 0$ is a pre-specified threshold.

Note that the original SRRS scheme in Lorden and Pollak [34] can be thought of as a limiting case of our proposed shrinkage-based scheme N_B in (1.3.8) when $a = 1, b = 0$ and $\omega_k \to 0$ in (1.3.5). In addition, many arguments in the asymptotic analysis of the

original SRRS scheme in Lorden and Pollak [34] for $K = 1$ dimension such as martingale properties and non-linear renewal theory for overshoot analysis can be applied to the proposed shrinkage-based scheme N_B, subject to a careful analysis of the shrinkage estimators in (1.3.5). Our major contribution is to introduce the shrinkage estimators to the quickest change detection problem and demonstrate its usefulness to lessen the dimension effects when the number K of data streams is large.

1.4 Asymptotic Properties

In this section, we investigate the asymptotic properties of the proposed shrinkage-based SRRS scheme N_B in (1.3.7) and (1.3.8) when the estimators $\widehat{\mu}_{k,m,\ell}$'s of the post-change means μ_k's are the shrinkage estimators in (1.3.5). The following discussion is divided into three subsections. The first two subsections address two properties of the proposed shrinkage-based SRRS scheme under the general setting: the ARL to false alarm and detection delay, respectively. The third subsection focuses on the suitable choice of tuning parameters in our proposed shrinkage-based SRRS scheme.

1.4.1 The ARL to False Alarm

To derive the ARL to false alarm of the proposed shrinkage-based SRRS scheme N_B in (1.3.7) and (1.3.8), it is crucial to observe that its global monitoring statistic R_n is the Shiryaev-Roberts-type statistics and thus $R_n - n$ is a martingale under the pre-change hypothesis. By the well-known Doob's optional stopping time theorem (see Theorem 10.10 of Williams [56]), for the stopping time $N = N_B$ defined in (1.3.8), we have $\mathbf{E}_\infty(N) = \mathbf{E}_\infty(R_N) \geq B$, as $R_{N_B} \geq B$ by the definition of N_B. Also see the proof of Theorem 4 of Lorden and Pollak [34] for more detailed arguments. The following theorem summarizes this result.

Theorem 1. *Consider the proposed shrinkage-based SRRS scheme N_B in (1.3.7) and (1.3.8) with $\widehat{\mu}_{k,m,\ell}$ being the shrinkage estimators in (1.3.5). For any $B > 0$,*

$$\mathbf{E}_\infty(N_B) \geq B.$$

While Theorem 1 is applicable regardless of the value of the dimension K (the number of data streams), it is important to point out that the Monte Carlo simulation of $\mathbf{E}_\infty(N_B)$ is a different story due to the curse of dimensionality. If the dimension K is small, say $K = 1$ or 5, then a Monte Carlo simulation with runs of thousands will provide a reasonable estimate of $\mathbf{E}_\infty(N_B)$ for a moderately large threshold B, say $B = 10^4$. However, the number of necessary runs is exponentially increasing as the dimension K increases, as the scheme N_B is highly skewed for large K, and the sample mean or median based on 10^5 or 10^6 of realizations of N_B can be a very poor estimate of $\mathbf{E}_\infty(N_B)$.

The reason is that the likelihood ratio $\Lambda_{n,m}(m < n)$ in (1.3.6) and the global monitoring statistic R_n in (1.3.7) are typically highly skewed to 0 and 1 for large dimensional K, respectively. To see this, consider the likelihood ratio $\Lambda_{n,m}$ when $\widehat{\mu}_{k,m,\ell}$ is the MLE/MOM estimates in (1.3.4). On the one hand, for a fixed n and any given $1 \le m \le n-1$, we have $\mathbf{E}_\infty(\Lambda_{n,m}) = 1$ and $\mathbf{E}_\infty(R_n) = n$. On the other hand, for normal distributions,

$$\begin{aligned}
& \mathbf{E}_\infty \log(\Lambda_{n,m}) \\
= & \sum_{\ell=m}^{n} \sum_{k=1}^{K} \mathbf{E}_\infty \Big(\mathbf{E}_\infty (\widehat{\mu}_{k,m,\ell} X_{k,\ell} - \frac{1}{2} (\widehat{\mu}_{k,m,\ell})^2 \Big| \widehat{\mu}_{k,m,\ell} \Big) \\
= & -\frac{1}{2} \sum_{\ell=m}^{n} \sum_{k=1}^{K} \mathbf{E}_\infty (\widehat{\mu}_{k,m,\ell})^2 \quad (\text{as } \mathbf{E}_\infty(X_{k,\ell}) = 0) \\
= & -\frac{1}{2} \sum_{k=1}^{K} \sum_{\ell=m+1}^{n} \frac{1}{\ell - m} \quad (\text{as } \widehat{\mu}_{k,m,m} = 0) \\
= & -\frac{1}{2} K (1 + \frac{1}{2} + \cdots + \frac{1}{n-m}).
\end{aligned}$$

Here the third equation uses the fact that when $\widehat{\mu}_{k,m,\ell}$ is the MLE/MOM estimates in (1.3.4), it has a $N(0, 1/(\ell - m))$ distribution. Now when $K = 100$, we have $\mathbf{E}_\infty \log(\Lambda_{n,n-1}) = -50$, implying that $\Lambda_{n,n-1}$ is concentrated around $e^{-50} = 1.9 \times 10^{-22}$, even though $\mathbf{E}_\infty(\Lambda_{n,n-1}) = 1$. For all other $m < n$, the likelihood ratios $\Lambda_{n,m}$'s will be concentrated around an even smaller value. Hence, for a fixed time n, when we simulate the global monitoring statistic R_n of the original SRRS scheme, we will mostly likely observe $R_n \approx 1$ (recall that $\Lambda_{n,n}$ is defined as a constant 1), although $\mathbf{E}_\infty(R_n) = n$.

The above argument can also be extended to the proposed SRRS scheme with the shrinkage estimators in (1.3.5), and our numerical experiences seem to suggest that the Monte Carlo estimate of $\mathbf{E}_\infty(N_B)$ works poorly and is highly biased unless the linear shrinkage factor a is of order $O(1/\sqrt{K})$. As mentioned in Rubinstein and Glynn [50], the curse of dimensionality is one of the central topics in Monte Carlo simulation due to the degeneracy properties of likelihood ratios, and the importance sampling technique does not help in the high dimensional problem unless we can reduce it to an equivalent low-dimension problem. It remains an open problem how to overcome the curse of dimensionality to simulate $\mathbf{E}_\infty(N_B)$ effectively for our proposed SRRS scheme N_B in the general context of monitoring a large number K of data streams.

A challenging practical question is how to find the threshold B of the proposed shrinkage-based SRRS scheme N_B in (1.3.7) and (1.3.8), so that it satisfies the pre-specified ARL to false alarm constraint A in (1.2.3). The good news is that Theorem 1 provides a theoretical bound: a choice of $B = A$ will guarantee that the proposed shrinkage-based SRRS scheme N_B satisfies the ARL to false alarm constraint in (1.2.3). For that reason, in our numerical simulations below, we will set $B = A$ and report the impact of shrinkage estimation on the detection delays of the proposed shrinkage-based SRRS scheme N_B.

1.4.2 Detection Delay

In this subsection, we derive the asymptotic expression of the detection delay of the proposed shrinkage-based SRRS scheme N_B in (1.3.7) and (1.3.8) under the setting when the dimension K is fixed and the threshold B goes to ∞. In this subsection and only in this subsection, we assume that a change occurs to the k-th data stream at time ν and the true post-change mean of the k-th data stream is μ_k for all $k = 1, \ldots, K$. That is, the true post-change mean vector $\mu = (\mu_1, \ldots, \mu_K)^T$. Recall that if $\mu_k = \mu_0$ then no changes occur to the k-th data streams.

To present the results on the detection delays, we need to first introduce some new notation. For each $k = 1, \ldots, K$, denote by μ_k^* the limit of the shrinkage estimators $\widehat{\mu}_{k,m,\ell}$ in (1.3.5) as $\ell \to \infty$, i.e., $\mu_k^* = \lim_{\ell\to\infty} \widehat{\mu}_{k,m,\ell}$ under the post-change hypothesis. Note that

the limit μ_k^* does not depend on the initial time m of the estimators, and for the shrinkage estimator $\widehat{\mu}_{k,m,\ell}$ in (1.3.5), it is easy to see that

$$\mu_k^* = \begin{cases} a\mu_k + b & \text{if } |\mu_k| > \omega_k \\ c & \text{if } |\mu_k| < \omega_k \end{cases}, \tag{1.4.1}$$

for each $k = 1, \ldots, K$. Here we purposely do not consider the cases of $\mu_k = \pm\omega_k$, as the corresponding analysis is complicated since the corresponding limit μ_k^* can be either $a\mu_k + b$ or c, either with probability $1/2$. This is also the reason why the post-change hypothesis set Ω in (1.2.1) makes a technical assumption that the post-change mean $\mu_k \neq \pm\omega_k$ so as to simplify our theoretical analysis.

Denote the vector of the limits μ_k^*'s in (1.4.1) by $\mu^* = (\mu_1^*, \ldots, \mu_K^*)^T$. It is important to note that in Lorden and Pollak [34], or more generally in the quickest change detection literature, the limit vector μ^* is always the same as the true post-change mean vector μ. Hence, the asymptotic analysis on the detection delay of the scheme N_B in (1.3.7) and (1.3.8) is closely related to the classical Shiryaev-Roberts procedure N_B^{SR} that detects a change from μ_0 to the known post-change μ. However, for our proposed shrinkage estimators in (1.3.5), it is no longer true that $\mu^* = \mu$. Hence, we need to compare N_B with the Shiryaev-Roberts procedure that mis-specifies the post-change means, i.e., the one that is designed to detect a change from μ_0 to μ^* but the true post-change mean vector is actually μ.

For that reason, we define a new information number

$$I(\mu^*, \mu_0; \mu) \;=\; \mathbf{E}_\mu \sum_{k=1}^{p} \left(\log \frac{f_{\mu_k^*}(X_{k,\ell})}{f_{\mu_0}(X_{k,\ell})} \right). \tag{1.4.2}$$

When f_{μ_0} and f_{μ_k} are normal distributions with common variance 1, it becomes

$$I(\mu^*, \mu_0; \mu) \;=\; -\frac{1}{2} \sum_{k=1}^{p} \mu_k^* (\mu_k^* - 2\mu_k). \tag{1.4.3}$$

Plugging the limits μ_k^*'s in (1.4.1) directly into (1.4.3) yields that

$$\begin{aligned} I(\mu^*, \mu_0; \mu) \;=\; & -\frac{1}{2} \sum_{k:|\mu_k|>\omega_k} \big(a\mu_k + b\big)\big((a-2)\mu_k + b\big) \\ & -\frac{1}{2} \sum_{k:|\mu_k|<\omega_k} c(c - 2\mu_k). \end{aligned} \tag{1.4.4}$$

In the special case when $(a, b, c) = (1, 0, 0)$, the new information number $I(\mu^*, \mu_0; \mu)$ in (1.4.4) has a simpler form:

$$I(\mu^*, \mu_0; \mu) \;=\; \frac{1}{2} \sum_{k:|\mu_k|>\omega_k} (\mu_k)^2. \tag{1.4.5}$$

If we further let ω_k go to 0, this becomes a more familiar form

$$I_{tot} = \frac{1}{2} \sum_{k=1}^{K} (\mu_k)^2, \tag{1.4.6}$$

where we change the notation to I_{tot} to emphasize its special meaning as the Kullback-Leibler divergence between the pre-change and post-change hypotheses. In Information Theory and Statistics, the Kullback-Leibler divergence I_{tot} in (1.4.6) has been regarded as a measure to characterize the distance between pre-change and post-change distributions, or equivalently, as a measure how difficulty it is to detect the change. However, in the context of monitoring a large K number of data streams, I_{tot} in (1.4.6) might no longer be as informative as one thought, since it ignores the spatial uncertainty associated with which subset of data streams are affected. A more meaningful measure that takes into account such a spatial uncertainty will be $I(\mu^*, \mu_0; \mu)$ in (1.4.5), or more generally those in (1.4.4) or (1.4.2), which can be thought of as the shrinkage version of the Kullback-Leibler divergence I_{tot} in (1.4.6).

With the notation in (1.4.4)-(1.4.6), we are ready to present our main result on the detection delays of the proposed shrinkage-based SRRS scheme N_B in (1.3.7) and (1.3.8).

Theorem 2. *Consider the proposed shrinkage-based SRRS scheme N_B in (1.3.7) and (1.3.8) with the estimators $\widehat{\mu}_{k,m,\ell}$'s defined in (1.3.5). Assume that the information number $I(\mu^*, \mu_0; \mu)$ in (1.4.4) is positive. Then, as $B \to \infty$, the detection delay of N_B satisfies*

$$\begin{aligned} D_\mu(N_B) \;\le\; & \frac{\log B}{I(\mu^*, \mu_0; \mu)} + \frac{ra^2}{2I(\mu^*, \mu_0; \mu)} \times \\ & \times \log\Big(\frac{\log B}{I(\mu^*, \mu_0; \mu)}\Big) + o(r \log\log B), \end{aligned} \tag{1.4.7}$$

where r is defined in (1.2.2) and represents the true number of affected data streams with "big" local changes.

The detailed proof of this theorem will be provided later, and we would like to discuss more on the theorem result first. Note that the original SRRS scheme N_B^{orig} proposed in Lorden and Pollak [34] can be thought of as the special case of our proposed scheme N_B in (1.3.7) and (1.3.8). By Theorems 1 and 2, or by extending the proof of Theorem 4 in Lorden and Pollak [34] from one-dimensional to K-dimensional, we can establish the first-order asymptotic optimality of the original SRRS scheme N_B^{orig} in Lorden and Pollak [34] as follows:

Corollary 1. *As $B \to \infty$, the original SRRS scheme N_B^{orig} satisfies*

$$\begin{aligned} D_\mu(N_B^{orig}) \;\; \leq \;\; & \frac{\log B}{I_{tot}} + \frac{K}{2I_{tot}} \log\Big(\frac{\log B}{I_{tot}}\Big) \\ & + o(K \log\log B), \end{aligned} \tag{1.4.8}$$

where I_{tot} is defined in (1.4.6). Moreover, if we let $B = A$, then the original SRRS scheme N_B^{orig} is first-order asymptotically optimal in the sense of asymptotically minimizing the detection delay $D_\mu(N_B^{orig})$ for each and every post-change mean vector μ subject to the false alarm constraint in (1.2.3) when K is fixed and the constraint A in (1.2.3) goes to ∞.

Proof. For the original SRRS scheme N_B^{orig}, the limit μ_k^* in (1.4.1) becomes the true post-change mean μ_k itself, and thus it is clear that the original SRRS scheme N_B^{orig} can be thought of as the special case of our proposed scheme N_B in (1.3.7) and (1.3.8) with $a = 1, r = K$ and $I(\mu^*, \mu_0; \mu) = I_{tot}$. Relation (1.4.8) then follows directly from Theorem 2. By Theorem 1, the choice of $B = A$ makes sure that the original SRRS scheme N_B^{orig} satisfies the false alarm constraint in (1.2.3). Then the asymptotic optimality properties of N_B^{orig} follows at once from a well-known lower bound on the detection delay of any scheme N satisfying the false alarm constraint in (1.2.3): $D_\mu(N) \geq (1 + o(1))(\log A)/I_{tot}$, see Lorden [33]. □

While Corollary 1 establishes the first-order asymptotic optimality property of the original SRRS scheme N_B^{orig}, it can be meaningless in practical setting when the dimension K is large and B is only moderately large. This is because the second-order term $(\log\log(B))$ in the right-hand side of (1.4.8) has coefficient K and can be significant as compared to

the first-order term $\log(B)$. A comparison of (1.4.7) and (1.4.8) shows that shrinkage estimators impact the detection delays in two different places: one is the information number $I(\mu^*, \mu_0; \mu)$ in (1.4.4) on the first-order term, and the other is the factor ra^2 in the second-order term. These will allow us to illustrate in the next subsection how a suitable choice of shrinkage estimators in (1.3.5) can reduce the overall detection delay.

Now let us turn back to the proof of Theorem 2.

Proof: To prove Theorem 2, the crucial technical tools are from those of Theorem 3 and part (iii) of Theorem 4 in Lorden and Pollak [34], which deals with $K = 1$ data stream without shrinkage. Below we will highlight the main difference with the dimension $K \geq 1$ and shrinkage.

To simplify the arguments, let us consider the hypothesis testing version of the quickest change detection problem, and assume that we want to test the null hypothesis H_0 : *no change* against the alternative hypothesis H_1 : *a change occurs exactly at time* $\nu = 1$. In such a problem, the corresponding sequential hypothesis testing version of the proposed scheme N_B in (1.3.7) and (1.3.8) is defined by

$$\tau_B = \inf\{n \geq 1 : \Lambda_n \geq B\}, \tag{1.4.9}$$

where the likelihood ratio

$$\Lambda_n = \prod_{\ell=1}^{n} \prod_{k=1}^{K} \frac{f_{\widehat{\mu}_{k,\ell}}(X_{k,\ell})}{f_0(X_{k,\ell})}, \tag{1.4.10}$$

and the estimate $\widehat{\mu}_{k,\ell}$ is a short-handed notation for $\widehat{\mu}_{k,m=1,\ell}$ in (1.3.5).

It is useful to mention that the quickest change detection scheme N_B in (1.3.7) and (1.3.8) is closely related to the sequential hypothesis testing procedure τ_B in (1.4.9) and (1.4.10), and such a close relation was first discovered in Lorden [33]. More specifically, for $t = 1, 2, \ldots,$ denote by $\tau_B^{(t)}$ the new stopping time that applies the sequential hypothesis testing procedure τ_B to the data starting from time t, i.e., $\{(X_{1,i}, \ldots, X_{K,i})\}$ for $i = t, t+1, \ldots$. Then the quickest change detection scheme $N_B = \min_{t \geq 1}\{\tau_B^{(t)} + t - 1\}$. This relation allows one to show that the detection delay $D(N_B)$ is asymptotically equivalent to $\mathbf{E}_\mu(\tau_B)$ under the alternative hypothesis H_1 when $\mu = (\mu_1, \cdots, \mu_K)^T$ is the true post-change mean

vector. To emphasize the dependence on the true μ, denote by $\mathbf{P}_\mu$ and $\mathbf{E}_\mu$ the corresponding probability mean and expectation under the alternative hypothesis H_1. Then it suffices to show that $\mathbf{E}_\mu(\tau_B)$ satisfies the right-hand side of (1.4.7).

Recall that in Section 1.4.2, we denote by μ_k^* the limit of $\widehat{\mu}_{k,\ell}$ under $\mathbf{P}_\mu$ for each $k = 1, \ldots, K$ as $\ell \to \infty$, and define $\mu^* = (\mu_1^*, \ldots, \mu_K^*)^T$. A key step of the proof is to relate Λ_n in (1.4.10) to the likelihood ratio Λ_n^* which mis-specify the true post change parameter μ_k of the k-th data stream as μ_k^* for all $k = 1, \ldots, K$. Since $\log \Lambda_n = 0$ when $n = 1$, we can define the mis-specified log-likelihood ratio by

$$\log \Lambda_n^* = \sum_{\ell=2}^{n} \sum_{k=1}^{K} \log \frac{f_{\mu_k^*}(X_{k,\ell})}{f_0(X_{k,\ell})}.$$

for $n \geq 2$ and $\log \Lambda_1^* = 0$. Then under $\mathbf{P}_\mu$, $\log \Lambda_n^*$ is a random walk with iid increments that have finite variance and mean $I(\mu^*, \mu_0; \mu)$ in (1.4.3).

For the stopping time $N = \tau_B$ in (1.4.9), applying Wald's equation to the random walk $\log \Lambda_n^*$ yields

$$\begin{aligned} I(\mu^*, \mu_0; \mu)\mathbf{E}_\mu(N) &= \mathbf{E}_\mu(\log \Lambda_N^*) \\ &= \mathbf{E}_\mu(\log \Lambda_N) + \mathbf{E}_\mu(\log \Lambda_N^* - \log \Lambda_N). \end{aligned}$$

For the notational convenience, let $b = \log B$. Then the standard renewal theorem for over-shoot analysis shows that $\mathbf{E}_\mu(\log \Lambda_N) = b + O(1)$ for $N = \tau_B$ in (1.4.9), where the $O(1)$ term is the over-shoot effect and may depend on the dimension K, see Theorem 3 of Lorden and Pollak [34]. Thus

$$I(\mu^*, \mu_0; \mu)\mathbf{E}_\mu(N) = b + O(1) + \mathbf{E}_\mu(\log \Lambda_N^* - \log \Lambda_N). \tag{1.4.11}$$

Hence it suffices to investigate the property of

$$\log \Lambda_N^* - \log \Lambda_N = \sum_{\ell=2}^{N} \sum_{k=1}^{K} \log \frac{f_{\mu_k^*}(X_{k,\ell})}{f_{\widehat{\mu}_{k,\ell}}(X_{k,\ell})}$$

when $N = \tau_B$ in (1.4.9) and $X_{k,\ell} \sim N(\mu_k, 1)$ under $\mathbf{P}_\mu$.

To do so, note that this involves the likelihood ratio of the form $f_{\mu_k^*}(X_{k,\ell})/f_{\phi_k}(X_{k,\ell})$ when $X_{k,\ell}$'s are iid $N(\mu_k, 1)$ for each k, and the ϕ_k's may vary and converge to μ_k^*. Thus

for any given $\phi = (\phi_1, \ldots, \phi_K)$, we need to define another information number:

$$\begin{aligned} I(\mu^*, \phi; \mu) &= \mathbf{E}_\mu \sum_{k=1}^{K} \left(\log \frac{f_{\mu_k^*}(X_{k,\ell})}{f_{\phi_k}(X_{k,\ell})} \right) \\ &= \sum_{k=1}^{K} \Big((\mu_k^* - \phi_k)\mu_k - \frac{1}{2}(\mu_k^*)^2 + \frac{1}{2}(\phi_k)^2 \Big) \\ &= \sum_{k=1}^{K} \Big((\mu_k^* - \mu_k)\Delta_k + \frac{1}{2}(\Delta_k)^2 \Big) \end{aligned} \tag{1.4.12}$$

where $\Delta_k = \phi_k - \mu_k^*$ for $k = 1, \ldots, K$. It is useful to compare this new information number with $I(\mu^*, \mu_0; \mu)$ in (1.4.2). On the one hand, they are defined similarly except that $\phi_k \equiv \mu_0 = 0$ for all k. On the other hand, $I(\mu^*, \mu_0; \mu)$ in (1.4.2) is related to the first-order term of the detection delay of τ_B, whereas $I(\mu^*, \phi; \mu)$ in (1.4.12) contributes to the second-order term of the detection delay when we let $\Delta_k = \phi_k - \mu_k^*$ go to 0 for all k.

For any given $\ell = 2, 3, \ldots$, let $\widehat{\mu}_\ell = (\widehat{\mu}_{1,\ell}, \ldots, \widehat{\mu}_{K,\ell})^T$, and let $I(\mu^*, \widehat{\mu}_\ell; \mu)$ be the information number defined in (1.4.12) when $\phi = \widehat{\mu}_\ell$. As in Lorden and Pollak [34], the application of the martingale optional sampling theorem to $\log \Lambda_n^* - \log \Lambda_n - \sum_{\ell=2}^{n} I(\mu^*, \widehat{\mu}_\ell; \mu)$ yields that

$$\mathbf{E}_\mu \Big(\log \Lambda_N^* - \log \Lambda_N \Big) = \mathbf{E}_\mu \sum_{\ell=2}^{N} I(\mu^*, \widehat{\mu}_\ell; \mu). \tag{1.4.13}$$

By (1.4.12), if we suppress the notation ℓ for the sake of convenience and let $\Delta_k = \widehat{\mu}_{k,\ell} - \mu_k^*$, then

$$\begin{aligned} \mathbf{E}_\mu \Big(I(\mu^*, \widehat{\mu}_\ell; \mu) \Big) &= \sum_{k=1}^{K} (\mu_k^* - \mu_k)\mathbf{E}_\mu(\Delta_k) + \\ &\quad + \frac{1}{2} \sum_{k=1}^{K} \mathbf{E}_\mu(\Delta_k^2), \end{aligned} \tag{1.4.14}$$

and thus the proof of Theorem 2 relies on the analysis of $\mathbf{E}_\mu(\Delta_k)$ and $\mathbf{E}_\mu(\Delta_k^2)$.

In a high-level description, we may expect that $\Delta_k = \widehat{\mu}_{k,\ell} - \mu_k^*$ converges to 0 as $\ell \to \infty$. Hence, for large ℓ, we should expect that $\mathbf{E}_\mu(\Delta_k) \approx 0$ becomes negligible, and the term $\mathbf{E}_\mu(\Delta_k^2) \approx Var(\Delta_k)$ may or may not be significant. Indeed, for a given ℓ, we will show below that as $\ell \to \infty$,

$$\mathbf{E}_\mu(\Delta_k) = o\Big(\frac{1}{(\ell - 1)^2}\Big) \tag{1.4.15}$$

and

$$\mathbf{E}_\mu(\Delta_k^2) \sim \begin{cases} a^2/(\ell-1), & \text{if } |\mu_k| > \omega_k; \\ o(\frac{1}{(\ell-1)^2}), & \text{if } |\mu_k| < \omega_k. \end{cases} \tag{1.4.16}$$

Let us postpone the proof of (1.4.15) and (1.4.16) in a little bit, and apply them directly to (1.4.14), we have

$$\begin{aligned} & \mathbf{E}_\mu\Big(I(\mu^*, \widehat{\mu}_\ell; \mu)\Big) \\ = & \sum_{k=1}^{K} o\Big(\frac{1}{(\ell-1)^2}\Big) + \frac{1}{2} \sum_{k:|\mu_k|>\omega_k} \frac{a^2}{\ell-1} + \\ & + \frac{1}{2} \sum_{k:|\mu_k|<\omega_k} o\big(\frac{1}{(\ell-1)^2}\big)\Big) \\ = & \frac{r}{2}\frac{a^2}{\ell-1} + o(\frac{1}{(\ell-1)^2}) \end{aligned}$$

as ℓ goes to ∞, where r is defined in (1.2.2). Plugging this into (1.4.13), we have

$$\mathbf{E}_\mu\Big(\log \Lambda_N^* - \log \Lambda_N\Big) = \frac{ra^2}{2}(1+o(1))\mathbf{E}_\mu \sum_{\ell=2}^{N} \frac{1}{(\ell-1)}.$$

The summation of the above relation can then be estimated as in Theorem 3 of Lorden and Pollak [34] by

$$(1+o(1)) \sum_{\ell=2}^{n_0} \frac{1}{(\ell-1)} \approx (1+o(1)) \log(n_0)$$

where $n_0 =$ the largest integer $\le \mathbf{E}_\mu(N)$. Combining this with (1.4.11) yields

$$\begin{aligned} & I(\mu^*, \mu_0; \mu)\mathbf{E}_\mu(N) \\ = & \ b + O(1) + \mathbf{E}_\mu(\log \Lambda_N^* - \log \Lambda_N) \\ = & \ b + O(1) + (1+o(1))\frac{ra^2}{2} \log(\mathbf{E}_\mu(N)). \end{aligned}$$

This gives an equation for $\mathbf{E}_\mu(N)$, and thus $\mathbf{E}_\mu(N)$ can be found by solving the equation of the form $x = \alpha + \beta \log(x)$ for large $\alpha > 0$ and possibly large $\beta > 0$. Taking logarithms of both sides yields

$$\begin{aligned} \log(x) &= \log(\alpha + \beta \log(x)) = \log \max\{\alpha, \beta \log(x)\} + O(1) \\ &= \max\{\log \alpha, \log \beta\} + o(\log x), \end{aligned}$$

where we use the fact that $\max(x,y) \le x+y \le 2\max(x,y)$ for $x>0, y>0$ and $O(1) = O(\log\log x) = o(\log x)$ for large x. Plugging this relation back to $x = \alpha + \beta\log(x)$ yields that

$$x = \alpha + (1+o(1))\beta \max\{\log\alpha, \log\beta\}.$$

Using the above arguments to derive $\mathbf{E}_\mu(N)$ and absorbing all insignificant terms to the $o(1)$ term, we have

$$\begin{aligned} & \mathbf{E}_\mu(N) \\ = \quad & \left(b + (1+o(1))\frac{ra^2}{2} \log \frac{\max\{b, ra^2/2\}}{I(\mu^*, \mu_0; \mu)} \right) / I(\mu_*, \mu_0; \mu) \end{aligned}$$

which becomes the right-hand side of (1.4.7) as $b = \log(B)$ goes to ∞. Thus the theorem holds.

It remains to prove (1.4.15) and (1.4.16). The details can be simplified to the following elementary probability question. Given two real numbers μ and $\omega > 0$, and $|\mu| \neq \omega$. Assume $Y = (X_{k,1} + \ldots + X_{k,\ell-1})/(\ell-1) \sim N(\mu, \sigma^2 = 1/(\ell-1))$, and define a new random variable

$$Y^* = \begin{cases} aY + b, & \text{if } |Y| \ge \omega; \\ c, & \text{if } |Y| < \omega. \end{cases}$$

and a new constant

$$\mu^* = \begin{cases} a\mu + b, & \text{if } |\mu| > \omega; \\ c, & \text{if } |\mu| < \omega. \end{cases}$$

Let $\Delta = Y^* - \mu^*$, and we want to show the asymptotic properties of $\mathbf{E}(\Delta)$ and $\mathbf{E}(\Delta^2)$ satisfy (1.4.15) and (1.4.16) as $\sigma^2 = \frac{1}{\ell-1} \to 0$.

We need to consider three cases, depending on the relationship between μ and $\pm\omega$. Following the traditional notation, let $Z = (Y-\mu)/\sigma \sim N(0,1)$, and denote by $\phi(z)$ and $\Phi(z)$ for the probability density function (pdf) and cumulative distribution function (cdf) of $N(0,1)$, respectively. Also define $\lambda_1 = (-\omega-\mu)/\sigma$ and $\lambda_2 = (\omega-\mu)/\sigma$. Then $Y^* = (a(\mu+\sigma Z)+b)(1\{Z \le \lambda_1\} + 1\{Z \ge \lambda_2\}) + c(1\{\lambda_1 \le Z \le \lambda_2\})$.

Let us focus on the case when $\mu > \omega$. In this case, we have $\mu^* = a\mu + b$ and $\lambda_1 < \lambda_2 \to -\infty$ as $\sigma \to 0$. Hence,

$$\begin{aligned}\Delta = Y^* - \mu^* &= a\sigma Z(1\{Z \le \lambda_1\} + 1\{Z \ge \lambda_2\}) \\ &\quad +(c - a\mu - b)1\{\lambda_1 < Z < \lambda_2\}.\end{aligned}$$

Since $\lambda_1 < \lambda_2 \to -\infty$ as $\sigma \to 0$, the event $1\{Z \ge \lambda_2\}$ is dominant whereas the other two events are rare events. Thus we should expect that $\Delta \approx a\sigma Z$, and thus $\mathbf{E}(\Delta) \approx o(\sigma^4)$ and $\mathbf{E}(\Delta^2) \approx Var(a\sigma Z) = a^2\sigma^2$. To be more rigorous,

$$\begin{aligned}\mathbf{E}(\Delta) &= \int_{-\infty}^{\lambda_1} a\sigma z\phi(z)dz + \\ &\quad + \int_{\lambda_2}^{\infty} a\sigma z\phi(z)dz + \int_{\lambda_1}^{\lambda_2} (c - a\mu - b)\phi(z)dz \\ &= -a\sigma\phi(|\lambda_1|) + a\sigma\phi(|\lambda_2|) + \\ &\quad +(c - a\mu - b)\mathbf{P}(|\lambda_2| \le Z \le |\lambda_1|)\end{aligned}$$

where we use the fact $\int_{-\infty}^{\lambda} z\phi(z)dz = -\phi(|\lambda|) = -\int_{\lambda}^{\infty} z\phi(z)dz$ when $\lambda < 0$. By the well-known fact that $\frac{x}{1+x^2}\phi(x) \le \mathbf{P}(Z > x) \le \frac{\phi(x)}{x}$ for all $x \ge 0$, it is clear that $\mathbf{E}(\Delta) = O(\sigma\phi(|\lambda_1|)) + O(\sigma\phi(|\lambda_2|)) = o(\sigma^4)$ as σ goes to 0, since $O(\phi(x/\sigma)) = O(\exp(-\frac{x^2}{2\sigma^2})) = o(\sigma^4)$ for any $x \ne 0$.

In addition,

$$\begin{aligned}\mathbf{E}(\Delta^2) &= \int_{-\infty}^{\lambda_1} (a\sigma z)^2\phi(z)dz + \\ &\quad + \int_{\lambda_2}^{\infty} (a\sigma z)^2\phi(z)dz + \int_{\lambda_1}^{\lambda_2} (c - a\mu - b)^2\phi(z)dz \\ &= a^2\sigma^2[\mathbf{P}(Z > |\lambda_1|) + |\lambda_1|\phi(|\lambda_1|)] \\ &\quad +a^2\sigma^2[1 - \mathbf{P}(Z > |\lambda_2|) + |\lambda_2|\phi(|\lambda_2|)] \\ &\quad +(c - a\mu - b)^2\mathbf{P}(|\lambda_2| \le Z \le |\lambda_1|) \\ &= a^2\sigma^2 + o(\sigma^4).\end{aligned}$$

Here in the second equation, we use the fact that $\int_{-\infty}^{\lambda} z^2\phi(z)dz = \mathbf{P}(Z > |\lambda|) + |\lambda|\phi(|\lambda|) = 1 - \int_{\lambda}^{\infty} z^2\phi(z)dz$ for $\lambda < 0$, which follows from the integration by parts for $z^2\phi(z) = -z(\phi(z))'$. Thus (1.4.15) and (1.4.16) hold for the case when $\mu > \omega$.

The above arguments can be easily extend to the other cases when $\mu < -\omega$ or $-\omega < \mu < \omega$. For instance, when $-\omega < \mu < \omega$, we have $\mu^* = c$ and $\lambda_1 \to -\infty, \lambda_2 \to \infty$ as $\sigma \to 0$. Thus $\Delta = Y^* - c = (a(\mu + \sigma Z) + b - c)(1\{Z \le \lambda_1) + 1\{Z \ge \lambda_2\})$. Since the probabilities of both events $1\{Z \le -\lambda_1\}$ and $1\{Z \ge \lambda_2\}$ go to 0 exponentially as σ goes to 0, the above arguments can show that both $\mathbf{E}(\Delta)$ and $\mathbf{E}(\Delta^2)$ are negligible (order $o(\sigma^4)$), completing the proof of the theorem.

1.4.3 How to Choose Suitable Shrinkage Estimators?

In our proposed shrinkage-based SRRS scheme N_B in (1.3.7) and (1.3.8) with the estimators $\widehat{\mu}_{k,m,\ell}$'s defined in (1.3.5), there are two sets of tuning parameters: one is the lower bounds ω_k's and the other is the constant (a, b, c). The choices of the lower bounds ω_k's are straightforward, as they are pre-specified in the post-change hypothesis set Ω in (1.2.1). Below we will focus on the suitable choice of tuning parameter (a, b, c).

By Theorem 2, if we want to minimize the first-order term of the detection delay of the proposed shrinkage-based scheme N_B, then it suffices to maximize the information number $I(\mu^*, \mu_0; \mu)$ in (1.4.4). Hence, it is natural to define the "first-order" optimal choice of (a, b, c) as the one that maximizes $I(\mu^*, \mu_0; \mu)$ in (1.4.4). The following theorem provides the corresponding "first-order" optimal choice of (a, b, c) among all possible shrinkage estimators $\widehat{\mu}_{k,m,\ell}$'s in (1.3.5):

Theorem 3. *Under the post-change hypothesis set Ω in (1.2.1), the choice of $a = 1, b = 0, c = 0$ is "first-order" optimal for the proposed SRRS scheme among all possible shrinkage estimators $\widehat{\mu}_{k,m,\ell}$'s in (1.3.5).*

Proof. It suffices to show that $a = 1, b = 0, c = 0$ maximizes $I(\mu^*, \mu_0; \mu)$ in (1.4.4). Note that the right-hand side of (1.4.4) is a quadratic function of a, b, c and thus the optimal values can be found by taking derivatives of the right-hand side of (1.4.4) with respect to

a, b, c. Following the definition of r in (1.2.2), define

$$\begin{aligned} D_1 &= \sum_{k=1}^{K} \mu_k 1\{|\mu_k| > \omega_k\} \\ D_2 &= \sum_{k=1}^{K} \mu_k^2 1\{|\mu_k| > \omega_k\} \\ D_3 &= \sum_{k=1}^{K} \mu_k 1\{|\mu_k| < \omega_k\}. \end{aligned}$$

Then the derivatives of the right-hand side of (1.4.4) with respect to a, b, c can be rewritten as

$$\begin{aligned} D_1(a-1) + rb &= 0; \\ D_2(a-1) + D_1 b &= 0; \\ (K-r)c - D_3 &= 0. \end{aligned}$$

Clearly, under the post-change hypothesis Ω in (1.2.1), the post-change mean $\mu_k = 0$ whenever $|\mu_k| \leq \omega_k$, implying that $D_3 = 0$. Hence, $(a^*, b^*, c^*) = (1, 0, 0)$ is the unique optimal choice of (a, b, c) when $(D_1)^2 \neq rD_2$ and $r \neq K$, and is one of infinitely many optimal solutions otherwise. Thus the theorem holds. □

When $a = 1, b = 0, c = 0$, the shrinkage estimators $\widehat{\mu}_{k,m,\ell}$'s in (1.3.5) become the hard-thresholding estimators

$$\widehat{\mu}_{k,m,\ell} = \begin{cases} \bar{X}_{k,m,\ell} & \text{if } \ell = m+1, \ldots, n, \text{ and} \\ & |\bar{X}_{k,m,\ell}| \geq \omega_k \\ \mu_0 = 0 & \text{otherwise} \end{cases}, \tag{1.4.17}$$

where $\bar{X}_{k,m,\ell}$'s are the MLE/MOM estimates of μ_k in (1.3.4). Denote by N_B^{hard} the corresponding SRRS scheme N_B in (1.3.7) and (1.3.8) when the estimators $\widehat{\mu}_{k,m,\ell}$'s being the hard-thresholding estimators (1.4.17). The following corollary summarizes its first-order asymptotic optimality properties:

Corollary 2. *For any fixed dimension K, the hard-thresholding scheme N_B^{hard} with $B = A$ asymptotically minimizes the detection delay $D_\mu(N_B^{hard})$ (up to first-order) for each and*

every post-change mean vector $\mu \in \Omega$ in (1.2.1) subject to the false alarm constraint A in (1.2.3) as the constraint A goes to ∞.

Proof. By Theorem 2 and Corollary 1, it suffices to show that for the hard-thresholding estimators in (1.4.17), $I(\mu^*, \mu_0; \mu)$ in (1.4.5) is the same as I_{tot} in (1.4.6) when $\mu \in \Omega$ in (1.2.1). From the definition of Ω in (1.2.1), we have $\mu_k = 0$ if $|\mu_k| < \omega_k$. Thus $\sum_{k:|\mu_k|>\omega_k} (\mu_k)^2 = \sum_{k=1}^{K} (\mu_k)^2$ and it is clear from (1.4.5) and (1.4.6) that $I(\mu^*, \mu_0; \mu) = I_{tot}$ for any $\mu \in \Omega$. Hence the corollary holds. □

It is useful to compare the original SRRS scheme N_B^{orig} in Corollary 1 with the hard-thresholding scheme N_B^{hard} in Corollary 2. On the one hand, the first-order asymptotic optimality property of N_B^{orig} is applicable to all possible post-change mean vectors μ no matter whether $\mu \in \Omega$ or not, whereas N_B^{hard} is first-order asymptotically optimal only for those $\mu \in \Omega$ in (1.2.1). On the other hand, for these two schemes, the coefficients in the second-order terms of the detection delays are different: K for N_B^{orig}, and r in (1.2.2) for N_B^{hard}. This is exactly the reason why the hard-thresholding estimators can reduce the detection delay in the sparse post-change case of Ω in (1.2.1) when the number of affected data streams is much smaller than the total number of data streams, e.g., when $r = 20$ out of $K = 100$ data streams are affected.

Corollary 2 also provides a partial answer to an open problem raised on page #426 of Mei [35] whether we can develop new methods to reduce the coefficient in the second-order term of the detection delay from K to a smaller number while keeping the first-order asymptotic optimality properties. Our results show that such coefficient can be reduced to the number r of affected data streams in the sparse post-change case. We conjecture that r in (1.2.2) is the smallest possible coefficient for the second-order term in the Gaussian model, but we do not have a rigorous proof.

Besides the sparse post-change case, another interesting case of Ω in (1.2.1) is when all data streams are affected simultaneously. In this case, we have $r = K$, and thus the hard-thresholding scheme N_B^{hard} does not necessarily work efficiently, and to the best of our knowledge, no methodologies have been developed to improve the original SRRS scheme

N_B^{orig} or other classical quickest change detection schemes when the unknown local post-change means might be different for different local data streams. Below we will demonstrate how to use Theorem 2 to derive a good choice of the linear shrinkage factor a that can balance the tradeoff between the first-order and second-order of the detection delay.

To highlight our main ideas, let us focus on a by setting $b = c = 0$. Then the estimators $\widehat{\mu}_{k,m,\ell}$'s in (1.3.5) becomes

$$\widehat{\mu}_{k,m,\ell} = \begin{cases} a\bar{X}_{k,m,\ell} & \text{if } \ell = m+1, \ldots, n, \text{ and} \\ & |\bar{X}_{k,m,\ell}| \geq \omega_k \\ \mu_0 = 0 & \text{otherwise} \end{cases} \tag{1.4.18}$$

for some $0 \leq a \leq 1$, where $\bar{X}_{k,m,\ell}$'s are the MLE/MOM estimates of μ_k in (1.3.4). Then $I(\mu^*, \mu_0; \mu)$ in (1.4.4) becomes $I(\mu^*, \mu_0; \mu) = a(2-a)I_{tot}$. By Theorem 2, when $r = K$, minimizing the detection delay of N_B is asymptotically equivalent to minimizing

$$\frac{\log B}{a(2-a)I_{tot}} + \frac{Ka}{2(2-a)I_{tot}} \log\Big(\frac{\log B}{I_{tot}}\Big) \tag{1.4.19}$$

if we only keep the key terms containing the factor a and ignore the $1/(a(2-a))$ factor inside the logarithm of the second term. Clearly, $a = 1$ maximizes $a(2-a)$, and this is equivalent to the first-order asymptotic optimality properties of N_B^{orig} or N_B^{hard}. However, a better choice of a is to find $0 < a \leq 1$ that minimize the summation in (1.4.19), not just the first term in (1.4.19). Note that a choice of $0 < a \leq 1$ will make sure that the factor $a/(2-a)$ in the second term of (1.4.19) is less than 1. The corresponding optimal value of a will depend on I_{tot}, $\log B$, and K. For instance, when $B = 5000, K = 100$ and $I_{tot} = 2.5$, the summation in (1.4.19) becomes

$$\frac{3.407}{a(2-a)} + \frac{24.516a}{2-a}.$$

This summation has the value 46.2 when $a = 1$, and is minimized at $a = 0.25$ with the smallest value 13.9. This suggests that a suitable choice of linear shrinkage estimators in (1.4.18) can greatly reduce the overall detection delay as compared to the original SRRS scheme, although the price we pay is to sacrifice the first-order asymptotic optimality properties.

1.4.4 More Theoretical Results

In this subsection, we provide some refiner asymptotic results that are not included in our published paper in 2015 on IEEE Transactions on Information Theory. Recall that in Theorem 1, we only provide a lower bound for ARL to false alarm rate for a general $\widehat{\mu}_{k,m,\ell}$ defined in eq. (1.4.1). In this subsection, we improve the result in Theorem 1 under a special case of the linear shrinkage estimator:

$$\begin{aligned} \widehat{\mu}_{k,m,\ell} &= a\bar{X}_{k,m,\ell} + b \\ &= a\bar{X}_{k,m,\ell} + (1-a)\zeta. \end{aligned} \tag{1.4.20}$$

Here $0 \leq a \leq 1$ is the shrinkage factor, $\zeta (= b/(1-a)$ for $a \neq 1, = 0$ otherwise) is a pre-specified real-valued constant.

The main result of the this subsection is summarized in the following theorem.

Theorem 4. *Consider the proposed shrinkage-based SRRS scheme N_B in (1.3.7) and (1.3.8) with $\widehat{\mu}_{k,m,\ell}$ being the linear shrinkage estimators in (1.4.20). Then, as $B \to \infty$,*

$$\lim_{B\to\infty} \mathbf{E}_\infty(N_B)/B = 1/\gamma \tag{1.4.21}$$

where the over-shoot factor

$$\gamma = \begin{cases} \nu(\sqrt{p}\zeta) & \text{when } 0 \leq c < 1; \\ \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \nu(\|y\|) dG(y_1) \ldots dG(y_p) & \text{when } c = 1. \end{cases} \tag{1.4.22}$$

and $\|x\| = \sqrt{x_1^2 + ... + x_p^2}$, $G(\cdot)$ is cdf of $N(0, \sum_{i=1}^{\infty} \frac{1}{i^2}) = N(0, \frac{\pi^2}{6})$, and $\nu(\cdot)$ is a function that often appears in the overshoot analysis in renewal theory:

$$\nu(\mu) = 2\mu^{-2} \exp\{-2 \sum_{n=1}^{\infty} n^{-1} \Phi(-\frac{1}{2}|\mu|\sqrt{n})\}$$

The γ value in (1.4.22) for the case $c = 1$ can be further simplified from multiple integral to single integral, but we keep the current form so as to better understand its probability and statistical meaning.

Theorem 4 is useful to answer the following question: how to choose the threshold B for the stopping time N_B in (1.3.8) with $\widehat{\mu}_{k,m,\ell}$ defined in (1.4.1) and $\omega_k = 0$, so that it satisfies

the false alarm constraint in (1.2.3). It is clear from Theorem 4 that a more accurate choice is $B \approx A \times \gamma$, whereas Theorem 1 yields a more conservative choice $B = A$.

In order to prove Theorem 4, a crucial step is to introduce a one-sided stopping time τ_B defined in (1.4.9) and (1.4.10), where $\widehat{\mu}_{k,\ell}$ is a short-handed notation for linear-shrinkage $\widehat{\mu}_{k,m=1,\ell}$ in (1.4.20). We need to investigate the properties of $\mathbf{P}_\infty(\tau_B < \infty)$, which can be done by the change of measure techniques, see [34]. To do so, let Q be a probability measure on $\{X_{k,1}, X_{k,2}, \ldots\}$ such that for each $k = 1, \ldots, K$,

$$X_{k,n}|X_{k,1}, \ldots, X_{k,n-1} \sim N(\widehat{\mu}_{k,n}, 1), \quad \widehat{\mu}_{k,n} = a\bar{X}_{k,n} + (1-a)\zeta, \quad n = 1, 2, \ldots.$$

In other words, under the new probability measure Q, $X_{k,n} = \widehat{\mu}_{k,n} + Z_{k,n}$, where the $Z_{k,n}$'s are i.i.d. $N(0,1)$ distributed.

An important technical detail is to investigate the asymptotic behavior of $\widehat{\mu}_{k,n}$ under the new probability measure Q as $n \to \infty$. The case of $a = 1$ was studied on page 1442 in [34], and it was shown that $\{\widehat{\mu}_{k,n}\}$ converges to a random variable with distribution $G = N(0, \sum_{i=1}^{\infty} \frac{1}{i^2}) = N(0, \frac{\pi^2}{6})$. However, similar arguments and conclusions do not work for the case of $0 \le a < 1$. For instance, when $a = 0$, we have $\widehat{\mu}_{k,n} \equiv \zeta$ for all n. It turns out that when $0 < a < 1$, the properties of $\widehat{\mu}_{k,n}$ under Q are similar to the case of $a = 0$ rather than to the case of $a = 1$.

The following proposition summaries our results when $0 \le a < 1$.

Proposition 1. *Assume $0 \le a < 1$. Then the sequence $\{\widehat{\mu}_{k,n}\}$ converges almost surely to the real number ζ under measure Q for all $k = 1, \ldots, K$.*

Proof of Proposition 1: To prove Proposition 1, it suffices to prove it when $0 < a < 1$ for a fixed k. To simplify our notation, we drop the subscript k, and without loss of generality, assume $\zeta = 0$ (otherwise let $x^*_{k,n} = x_{k,n} - \zeta, \widehat{\mu}_{k,n} = \widehat{\mu}_{k,n} - \zeta$). In other words, we focus on a single sequence of random variables $X_1, X_2, \ldots$, whose conditional distribution are defined by

$$X_n|X_1, ..., X_{n-1} \sim N(\widehat{\mu}_n, 1)$$

and

$$\widehat{\mu}_n = a\big(\frac{X_1 + \ldots + X_{n-1}}{n-1}\big) \tag{1.4.23}$$

It remains to show that for $0 < a < 1$, the sequence $\{\widehat{\mu}_n\}$ in (1.4.23) converges to 0 a.s.. At the high-level, we want to show that $\mathbf{E}(\widehat{\mu}_n) = 0$ and $Var(\widehat{\mu}_n) \to 0$ as $n \to \infty$. For that purpose, the essential idea is to re-write $\widehat{\mu}_n$'s in terms of $N(0,1)$'s so that we can avoid difficult conditional probabilities. Let $X_n = \widehat{\mu}_n + Z_n$, and then $\widehat{\mu}_{n+1}$ in (1.4.23) can be rewritten as

$$\begin{aligned}
\widehat{\mu}_{n+1} &= a\frac{1}{n}\sum_{i=1}^{n} X_i = \frac{a}{n}[\frac{\widehat{\mu}_n}{a}(n-1) + \widehat{\mu}_n + Z_n] \\
&= \frac{n-1+a}{n}\widehat{\mu}_n + \frac{a}{n}Z_n = \cdots = \\
&= c_{n0}\widehat{\mu}_1 + \sum_{i=1}^{n-1} c_{ni} Z_i + c_{nn} Z_n
\end{aligned}$$

where

$$\begin{aligned}
c_{n0} &= \frac{(n-1+a)(n-2+a)\cdots(a+1)a}{n(n-1)\cdots 1} \\
c_{ni} &= \frac{a}{n}\prod_{k=i}^{n-1}(1+\frac{a}{k}), \qquad \text{for } \ 1 \le i \le n-1, \\
c_{nn} &= \frac{a}{n}
\end{aligned} \tag{1.4.24}$$

From this new representation of $\widehat{\mu}_{n+1}$, it is clear that $\mathbf{E}_Q(\widehat{\mu}_{n+1}) = 0$ and $Var_Q(\widehat{\mu}_{n+1}) = \sum_{i=1}^{n} c_{ni}^2$, as $\widehat{\mu}_1 = 0$. Thus the Proposition 1 holds as long as we can prove that

$$\sum_{i=1}^{n} c_{ni}^2 \to 0 \quad \text{ as } n \to \infty. \tag{1.4.25}$$

Note that this is also true for any arbitrary initial value of $\widehat{\mu}_1$ when $0 \le a < 1$, since we have c_{n0} goes to 0 as $n \to \infty$.

It remains to prove (1.4.25) for $0 < a < 1$. This requires us to approximate c_{ni} and $\sum_{i=1}^{n} c_{ni}^2$ via some classical calculus arguments for sufficiently large n. Let us fix (sufficiently large) n, and by abuse the notation, we write c_i for c_{ni} in (1.4.24). Note that

$$\log c_i = \log\frac{a}{n} + \sum_{k=i}^{n-1}\log(1+\frac{a}{k}).$$

Clearly, $\frac{x}{x+1} < \log(1+x) < x$ for $x > 0$, and thus

$$\frac{a}{k+1} < \frac{a}{k+a} < \log(1+\frac{a}{k}) < \frac{a}{k}.$$

Hence,

$$\log\frac{a}{n} + a\sum_{k=i}^{n-1}\frac{1}{k+1} \le \log c_i \le \log\frac{a}{n} + a\sum_{k=i}^{n-1}\frac{1}{k}.$$

Observe that $c_1 \ge c_2 \ge \cdots \ge c_n$ when $0 < a < 1$, and let us first approximate the largest value c_1. By the well-known fact that $\sum_{k=1}^{n}\frac{1}{k} - \log n$ converges to the Euler's constant $\beta \approx 0.5772$ (here we do not use the traditional notation γ, which unfortunately has already been used for the overshoot in (1.4.22) in Theorem 4), we have

$$\log\frac{a}{n} + a(-1 + \log n + \beta + o(1)) \le \log c_1 \le \log\frac{a}{n} + a(\log n + \beta + o(1))$$

and thus $c_1 = O(n^{-(1-a)})$, where $O(1)$ depends on a but not on n. Next, while it is nontrivial to get a good estimate of c_i for all $2 \le i \le n$, it is easy to do so when i is large, say, $i = n^s$ for some fixed $0 < s < 1$:

$$\log(c_{n^s}) \le \log\frac{a}{n} + a\sum_{k=n^s}^{n-1}\log(1+\frac{a}{k}) = \log\frac{a}{n} + a(\log n - \log n^s + o(1))$$

and thus $c_{n^s} = O(n^{-(1-a+as)})$, where $O(1)$ does not depend on n.

The key idea in the proof of (1.4.25) is to split $[1, n]$ into subintervals. Let us first try to split it into two subintervals $[1, n^s]$ and $[n^s, n]$ for some $0 < s < 1$. Since the c_i's are decreasing, we have

$$\begin{aligned}
\sum_{i=1}^{n} c_i^2 &= \sum_{i=1}^{n^s} c_i^2 + \sum_{i=n^s}^{n} c_i^2 \\
&\le n^s (c_1)^2 + (n - n^s)(c_{n^s})^2 \\
&\le n^s O(n^{-2(1-a)}) + nO(n^{-2(1-a+as)}) \\
&= O(n^{-2+s+2a}) + O(n^{-1+2(1-s)a}),
\end{aligned}$$

which converges to 0 as n goes to ∞ as long as $a < 1 - \frac{s}{2}$ and $a < \frac{1}{2(1-s)}$. If we choose $s = 1/2$, then this approach was able to prove (1.4.25) for $0 < a < 1 - 1/4 = 0.75$. Or a better choice is $s = (3-\sqrt{5})/2$, which can prove (1.4.25) for $0 < a < (\sqrt{5}+1)/4 \approx 0.809$. In other words,

splitting $[0,1]$ into two subintervals, we are able to prove (1.4.25) for $0 < a < 0.809$. It is natural to see what happens if we extend this approach to more than two subintervals.

Let us consider the case of three subintervals. Recall that when $s = 1/2$, the above simple arguments of two subintervals work fine on the subinterval $[\sqrt{n}, n]$ but lead a poor estimate on the subinterval $[1, \sqrt{n}]$. Hence one can further split $[1, n]$ into three intervals: $[1, n^r]$, $[n^r, \sqrt{n}]$ and $[\sqrt{n}, n]$ for some $0 < r < 1/2$, then

$$\begin{aligned}
\sum_{i=1}^{n} c_i^2 &\leq n^r (c_1)^2 + (\sqrt{n} - n^r)(c_{n^r})^2 + (n - \sqrt{n})(c_{\sqrt{n}})^2 \\
&\leq n^r O(n^{-2(1-a)}) + \sqrt{n} O(n^{-2(1-a+ar)}) + n O(n^{-2(1-a+a/2)}) \\
&= O(n^{-2+r+2a}) + O(n^{-1.5+2(1-r)a}) + O(n^{-1+a}),
\end{aligned}$$

which converges to 0 as long as $a \leq \min(1 - \frac{r}{2}, \frac{3}{4(1-r)}, 1)$ for some $0 \leq r < 1/2$. In particular, let $r = 1/4$, then the arguments with three intervals was able to prove (1.4.25) for $0 < a < 1 - 1/8$.

To prove (1.4.25) for any fixed $0 < a < 1$, let $m =$ the smallest integer that $a < 1 - 2^{-m}$, and define the $m + 1$ exponents r_t's as $r_m = 1$, $r_{t-1} = \frac{1}{2} r_t$ for $t = m, m-1, \ldots, 1$, and $r_0 = 0$. In other words, $r_t = (\frac{1}{2})^{m-t}$ for $t = 1, \ldots, m$. Then we split the interval $[1, n]$ into m subintervals, $[n^{r_{t-1}}, n^{r_t}]$ for $t = 1, \ldots, m$. By an abuse of notation, let $c_t^* = c_i$ when $i = n^{r_t}$. Then $c_t^* = O(n^{-(1-a+ar_t)})$ for $t = 0, 1, \ldots, m$ and

$$\begin{aligned}
\sum_{i=1}^{n} c_i^2 &\leq \sum_{t=1}^{m} (n^{r_t} - n^{r_{t-1}})(c_{t-1}^*)^2 \\
&\leq \sum_{t=1}^{m} n^{r_t} O(n^{-2(1-a+ar_{t-1})}) \\
&= \sum_{t=1}^{m} O(n^{r_t - 2 + 2a - 2ar_{t-1}}) \\
&= O(n^{r_1 - 2 + 2a}) + \sum_{t=2}^{m} O(n^{-2(1-a)(1-r_{t-1})}),
\end{aligned}$$

which converges to 0 as $n \to \infty$, since all $O(1)$ terms do not depend on n, and our choice of m and the definition of r_t's make sure that $r_1 - 2 + 2a < 0$ and $(1-a)(1-r_{t-1}) > 0$ for all $t \geq 2$. This shows that (1.4.25) indeed holds for any given $0 < a < 1$, completing the proof of Proposition 1.

Now we are ready to to prove Theorem 4. It is well known that there is a natural correlation between one-sided hypothesis testing problem and sequential change-point detection problem, see [43] and [58]. Recall τ_B is defined in (1.4.9) and (1.4.10). In order to prove Theorem 4, we only need to show that as $B \to \infty$,

$$\mathbf{P}_\infty(\tau_B < \infty) = (1 + o(1))\gamma/B, \tag{1.4.26}$$

where γ is the same as in (1.4.22). Using the change of measures technique, we will consider the properties of τ_B under the measure $\mathbf{P}_Q$ in Proposition 1. First, we need to show that $\mathbf{P}_Q(\tau_B < \infty) = 1$ for $0 \le a < 1$. Similar as in Lemma 2 of [34], the key observation is that the conclusion holds if $\sum_{\ell=1}^{n}\sum_{k=1}^{K}(\widehat{\mu}_{k,\ell})^2 \to \infty$ a.s. under probability measure Q. By Proposition 1, when $0 \le a < 1$, $\widehat{\mu}_{k,\ell} \to \zeta$ a.s., and thus $\sum_{\ell=1}^{n}\sum_{k=1}^{K}(\widehat{\mu}_{k,\ell})^2 \sim n\sum_{k=1}^{K}\zeta^2 \to \infty$ under Q as long as $\zeta \neq 0$. Hence $\mathbf{P}_Q(\tau_B < \infty) = 1$ for $0 \le a \le 1$.

For simplification of notation, we temporarily denote $b = \log B$. By the standard change-of-measure argument,

$$\mathbf{P}_{H_0}(\tau_B < \infty) \quad = \quad e^{-b}\mathbf{E}_Q(\exp\{-(\log \Lambda_{\tau_B} - b)\}),$$

where Λ_n is defined in (1.4.10).

Relation (1.4.26) is proved in Theorem 1 of [34], and the key idea is a renewal-theoretical analysis of $\mathbf{E}_Q(\exp\{-(\log \Lambda_{\tau_B} - b)\})$ under $\mathbf{P}_Q$. A high-level rough sketch is as follows. Let us consider another stopping time $T = \tau_{e^{b-c}}$ for some constant $c > 0$ that is sufficiently large but relatively smaller than b. On the one hand, $\log \Lambda_T \ge b - c$ by definition. On the other hand, we can choose c large enough as compared to the overshoot so that $\log \Lambda_T \le b - c/2$ with a high probability. In such a case, let $\widehat{\mu}_{k,T} = y_k$ for $k = 1, \ldots, K$, then for $n \ge T$, $\widehat{\mu}_{k,n} \approx y_k$ and the increments of $\log \Lambda_n$ acts like i.i.d. random variables:

$$\begin{aligned}
\log \Lambda_n - \log \Lambda_T \quad &\approx \quad \sum_{\ell=T+1}^{n}\sum_{k=1}^{p}\log\frac{f_{y_k}(X_{k,\ell})}{f_0(X_{k,\ell})} \\
&= \quad \sum_{\ell=T+1}^{n}\sum_{k=1}^{p}(y_k X_{k,\ell} - \frac{y_k^2}{2}) \approx \|y\|N(0,1) + \frac{\|y\|^2}{2}.
\end{aligned}$$

The last approximation holds as $\mathbf{E}_Q(X_{k,\ell}) = \widehat{\mu}_{k,n} \approx y_k$. Note that $(\log \Lambda_n - \log \Lambda_T)$ is reduced to a one-dimensional random walk with increments of the form $||y||Z_i + (1/2)||y||^2$

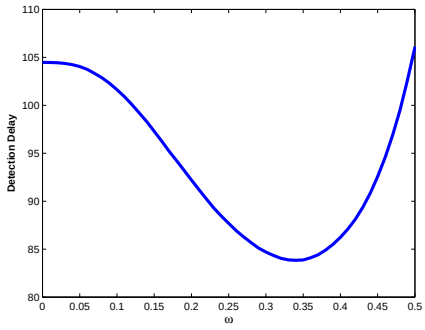

Figure 1: The sparse post-change case with the hard-thresholding scheme $N_B^{hard}(\omega)$ that sets the common lower bound $\omega_k \equiv \omega$ for all $k = 1, \ldots, K$. The x-axis is the common lower bound ω, and the y-axis is the simulated detection delay of $N_B^{hard}(\omega)$ when $B = 5000$. Here $\omega = 0$ corresponds to the baseline scheme N_B^{orig} without hard-thresholding.

with $Z_i \sim N(0,1)$. Thus the standard linear renewal theory can be applied to show the corresponding (conditional on $T = \tau_{e^{b-c}}$) overshoot factor is $\nu(\|y\|)$. Meanwhile, if $\widehat{\mu}_{k,n}$ converges a.s. to a random variable whose distribution is $G(\cdot)$ under Q, then the $y_k = \widehat{\mu}_{k,T}$'s have a distribution G. Combining them together yields the formula of γ in (1.4.22), see Theorem 1 of [34] for the detailed proof for $a = 1$. The proof for $0 \leq a < 1$ is identical, except that Proposition 1 now shows that $\widehat{\mu}_{k,n}$ converges to a real number ζ, i.e., the probability measure G will degenerate to Dirac measure δ_0 which defined only at a single atom $(\zeta, ..., \zeta)^T$, and thus (1.4.22) also holds for $0 \leq a < 1$. This concludes the proof of equation (1.4.26) in Theorem 4.

1.5 Numerical Simulations

In this section, we report numerical simulations to illustrate the usefulness of shrinkage or thresholding in the context of quickest change detection in Section 1.5.1, and demonstrate the challenge of Monte Carlo simulations of the ARL to false alarms when monitoring large-scale data streams in Section 1.5.2.

1.5.1 Shrinkage Effects

Assume we are monitoring $K = 100$ independent normal data streams whose initial distributions are $N(0,1)$ with possible changes in the means of some data streams. The ARL to false alarm constraint in (1.2.3) is assumed to be $\mathbf{E}_\infty(N) \geq A = 5000$. As mentioned in

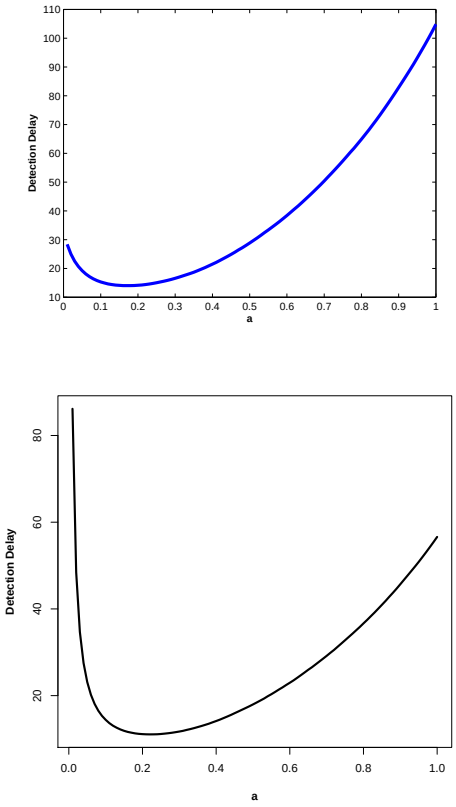

Figure 2: The case when all data streams are affected, and we consider the SRRS scheme $N_B(a)$ with varied linear shrinkage factor a while fixing $b = c = 0$ with two different choices of fixed lower bounds ω_k's: the upper plot is when $\omega_k = 0$ for all k, and the bottom plot is when $\omega_k = 0.01$ for all k. The x-axis is the value of linear shrinkage factor a, and the y-axis is the simulated detection delay of $N_B(a)$ when $B = 5000$. Here $a = 1$ corresponds to the scheme without linear shrinkage.

subsection 1.4.1, we set all thresholds $B = 5000$ for all schemes N_B's to avoid poor Monte Carlo estimates of $\mathbf{E}_\infty(N_B)$.

We have conducted extensive simulations for different schemes under different kinds of post-change hypothesis Ω in (1.2.1), but will only report the results of two specific post-change hypotheses so as to highlight our findings. The first one is the sparse post-change hypothesis case when $r = 20$ out of $K = 100$ data streams are affected, and the other is when all $K = 100$ data streams are affected. In both cases, we fix the overall information, $I_{tot} = \frac{1}{2}\sum_{k=1}^{K} \mu_k^2$, to be 2.5. To be more specific, we consider two cases: (1) when $r = 20$ out of $K = 100$ data streams are changed with the post-change mean $\mu_k = 0.5$ whereas there are no changes to the other remaining $K - r = 80$ data streams; and (2) when all $\mu_k = \sqrt{5/100} = 0.2236$ for all $k = 1, \ldots, K$. In both two cases, $I_{tot} = \frac{1}{2}\sum_{k=1}^{K} \mu_k^2 = 2.5$. However, when we design the monitoring scheme, we will only know that the post-change mean vector μ is in (1.2.1), and will not use any other information of the true post-change parameters. As shown in the second remark on Page 1435 in Lorden and Pollak [34], the worst-case detection delays of the SRRS scheme N_B occurs at time $\nu = 1$, and thus we will report the detection delay performance of $N_{B=5000}$ under the post-change hypothesis when the change occurs at time $\nu = 1$. All simulation results are based on 2500 replications.

For the purpose of comparison, the baseline scheme is the original SRRS scheme N_B^{orig} proposed by Lorden and Pollak [34]. In the sparse post-change case when $r = 20$ out of $K = 100$ data streams are affected, we consider several different kinds of hard-thresholding schemes N_B^{hard}'s in Corollary 2. For the convenience of comparison, we set $\omega_k \equiv \omega$ for all k, and then vary ω from 0 (baseline) to 0.5 with step size 0.01. For each hard-thresholding scheme with a given threshold ω, we then plot the detection delay of $N_B^{hard} = N_B^{hard}(\omega)$ as a function of ω in Figure 1. It is evident from Figure 1 that the detection delay of the scheme $N_B^{hard}(\omega)$ is reduced from 104.9 at $\omega = 0$ (baseline N_B^{orig}) to 83.8 at $\omega = 0.35$. This illustrates the usefulness of hard-thresholding estimators in the sparse post-change case.

In the case when all $K = 100$ data streams are affected with the post-change mean $\mu_k = 0.2236$ for all k, we consider two choices of the lower bound ω_k's: one is $\omega_k = 0$ for all k and the other is $\omega_k = 0.01$ for all k. The former choice of $\omega_k = 0$'s allows us to

see the performance of the original SRRS scheme N_B^{orig}. For each of these two choices of ω_k's, we vary the linear shrinkage factor a from 0.01 to 1, and then plot the detection delay of N_B as a function of a in Figure 2. It is clear from Figure 2 that the linear shrinkage can reduce detection delay from 104.8 at $a = 1$ (baseline N_B^{orig}) to 14.0 at $a = 0.17$ when the lower bound $\omega_k \equiv 0$ for all k, and can reduce detection delay from 56.6 at $a = 1$ to 11.1 at $a = 0.22$ when the lower bound $\omega_k \equiv 0.01$ for all k. Thus both hard-threshold ω_k's and linear shrinkage factor a can reduce the detection delay in this case, though the linear shrinkage factor a seems to be able to play more significant role. This is consistent with our asymptotic results in Section 1.4.3.

It is interesting to compare the sparse post-change case in Figure 1 with the simultaneous local changes case in Figure 2. In both cases, the overall Kullback-Leibler divergence $I_{tot} = 2.5$ are the same, and thus it is not surprising that the original SRRS scheme N_B^{orig} of Lorden and Pollak [34] has similar detection delays in these two cases (i.e., 104.9 versus 104.8). However, the smallest detection delay (i.e., 83.8) in Figure 1 in the sparse post-change case is much larger than the smallest detection delay (i.e., 11.1) in Figure 2 when all data streams are affected. In other words, given the same amount of Kullback-Leibler divergence information, it is much easier to detect simultaneous "small" local changes in all data streams than to detect "big" changes in a few unknown data streams if we incorporate relevant prior knowledge appropriately. This is consistent with our intuition since the latter has to deal with the uncertainty of the subset of affected data streams, which can be very challenging when the dimension K is large.

1.5.2 More Simulation About "Curse of Dimensionality"

In this section, we conduct Monte Carlo simulations to compare the empirical pre-change distributions of the global monitoring statistics R_n in (1.3.7) under two different dimensions: $K = 1$ and $K = 100$, thereby illustrating the challenge of Monte Carlo simulation of the ARL to false alarm when monitoring large $K > 1$ number of data streams.

We again assume to monitor K independent normal data streams, and each data stream follows distribution $N(0, 1)$. We focus on the performance of the original SRRS scheme

N_B^{orig} of Lorden and Pollak [34] and the corresponding R_n in (1.3.7) under the pre-change hypothesis (i.e. $\nu = \infty$). For each scenario of $K = 1$ and $K = 100$, and for each time step $n = 1, ..., 1000$, we ran Monte Carlo to simulate R_n with 2500 replications.

Figure 3 shows the histogram of R_n at a fixed time $n = 500$ for both $K = 1$ and $K = 100$ cases based on 2500 replications. As we can see, the empirical distribution of R_n is highly skewed for $K = 1$ with values in the range of $[0, 4000]$, but R_n seems to be empirically normally distributed for $K = 100$ with values in the range of $[0.998, 1.003]$. Theoretically the empirical mean of R_n with $n = 500$ should be 500 no matter whether the dimension $K = 1$ or $K = 100$. This suggests that 2500 replications might be sufficient for $K = 1$ dimension, but definitely not large enough for $K = 100$.

To further explain this issue, we also investigate the dynamic evolution of R_n over time n. To better illustrate, Figure 4 plots 2500 simulated $\log(R_n)$ versus $\log(n)$ for both $K = 1$ and $K = 100$ cases. From Figure 4, there is a clear linear trend of $\log(R_n)$ versus $\log(n)$ when the dimension $K = 1$, which matches the martingale property $\mathbf{E}_\infty(R_n) = n$. On the other hand, when the dimension $K = 100$, most of $\log(R_n)$ are 0, which implies that 2500 replications are not large enough to represent the property of R_n. The situation is similar even if we increase the number of Monte Carlo runs from 2500 to a larger number such as 10^4. All these simulations results are consistent with our theoretical results.

1.6 Conclusions

In this article, we investigated the quickest change detection problem in the context of monitoring independent large-scale normally distributed data streams when the post-change means are unknown. The key assumption we make is that for each individual local data stream, either there are no local changes, or there is a "big" local change. Our main contribution is to introduce the shrinkage estimators to quickest change detection, and show that the shrinkage estimators of the unknown post-change parameters can reduce the overall detection delays by balancing the tradeoff between the first-order and second-order terms of the asymptotic expression on the detection delays. Specifically, hard-thresholding is attractive in the sparse post-change case when the unknown number of affected data streams

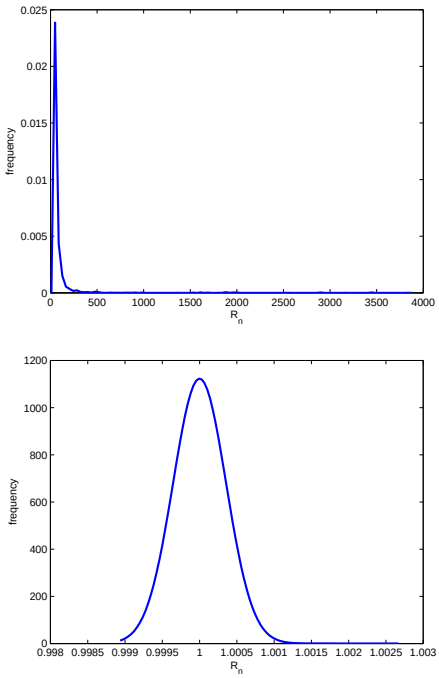

Figure 3: Histogram of 2500 simulated R_n with n=500 under two scenarios. *Upper Panel:* $K = 1$, and *Lower Panel:* $K = 100$.

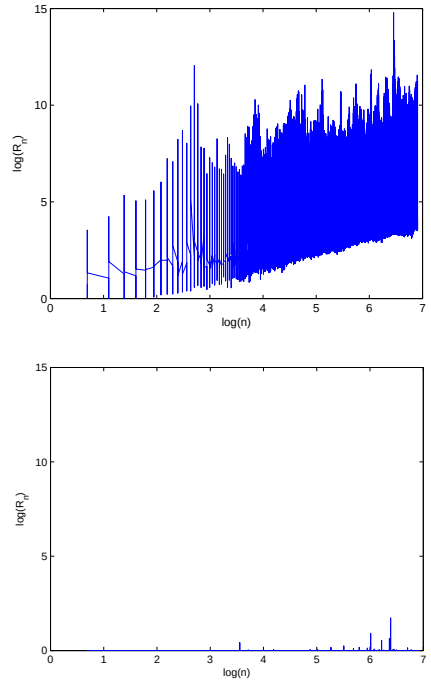

Figure 4: Plot of $(\log n, \log R_n)$ for $n = 1, .., 1000$ with 2500 replications under two scenarios. *Upper Panel:* $K = 1$, *Lower Panel:* $K = 100$.

is much smaller than the total number of data streams, whereas the linear shrinkage can be useful when all local data streams are affected simultaneously though not necessarily identically. Moreover, we illustrate the challenge of Monte Carlo simulation of of the ARL to false alarm when monitoring a large K number of data streams.

While the classical quickest change detection problems have been studied for several decades, further research on the quickest change detection for monitoring large-scale data streams is needed. For instance, in this paper we focus on the Gaussian model with known variances, and it will be interesting to extend the shrinkage estimators to a more general Gaussian model when the variances in the different data streams are different and unknown under the post-change hypothesis, or to other distributions such as Poisson. The corresponding theoretical analysis will likely be more challenging, e.g., the definition of r in (1.2.2) will need to be modified for other distributions such as Poisson. Moreover, it remains an open problem how to overcome the curse of dimensionality to conduct Monte Carlo simulations of the ARL to false alarm efficiently in the context of large-scale data streams. Hopefully this article can stimulate further research on quickest change detection problems in high-dimensional data streams.

CHAPTER II

THRESHOLDED MULTIVARIATE PRINCIPAL COMPONENT ANALYSIS FOR MULTI-CHANNEL PROFILE MONITORING

2.1 Introduction

Profile monitoring plays an important role in manufacturing systems improvement (Noorossana et al. [38], Qiu [45]), and a standard setup is to monitor a sequence of profiles (e.g. curves or functions) over time to check whether the underlying functional structure of the profiles changes or not. Extensive research has been done for monitoring *univariate* profile or *real-valued* functions in the area of statistical process control (SPC) in the past decades, and standard approaches are to reduce the univariate profiles in the infinite-dimensional or high-dimensional functional spaces to a low-dimensional set of features (e.g., shape, magnitude, frequency, regression coefficients, etc.). See, for instance, work by Jin and Shi [27], Ding et al. [13], Jeong et al. [25], Jensen et al. [24], Berkes et al. [4], Chicken et al. [10], Qiu et al. [46], Abdel-Salam et al. [1].

Nowadays manufacturing systems are often equipped with a variety of sensors capable of collecting several profile data simultaneously, and thus one often faces the problem of monitoring multichannel or multivariate profiles that have rich information about systems performance. A concrete motivating example of this paper is from a forging process, shown in Figure 5 and 6, in which multichannel load profiles measure exerted forces in each column of the forging machine. Here each data is a four-dimensional vector function or four curves that have similar but not identical shapes when the machine is operating under the normal state. While various methods have been developed for univariate profile monitoring, they often cannot easily be extended to multichannel profiles, and research on monitoring multivariate/mutichannel nonlinear profiles is very limited. For some exceptions, see Jeong et al. [26], Paynabar et al. [40], Grasso et al. [18], and Paynabar et al. [41]. There are two main challenges when monitoring multichannel profiles. The first one is that profiles

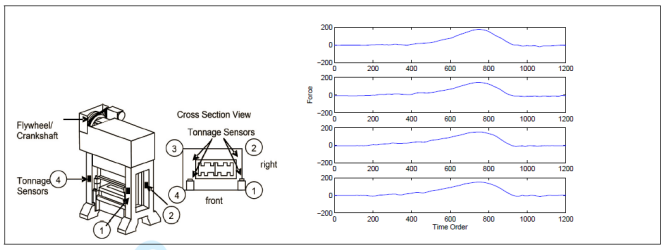

Figure 5: *Left:* A forging machine with 4 tonnage sensors. *Right:* A single run sample of four-dimensional functional data.

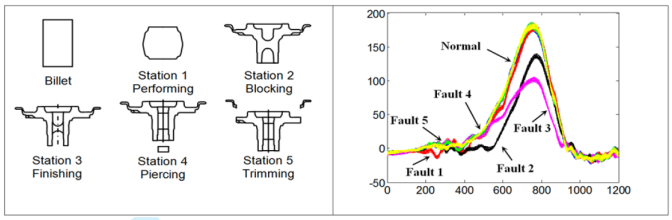

Figure 6: *Left:* Shape of workpieces at each operation. *Right:* Tonnage profile for normal and missing operations.

are high-dimensional functions with intrinsic inner- and inter-channel correlations, and one needs to develop a dimension reduction method that can deal with such intricate correlations. The second, probably more fundamental, challenge is that the functional structure of multi-channel profiles might change over time, and thus the dimension reduction method should be able to take into account the potential unknown change.

The primary goal of this paper is to develop an effective statistical method for monitoring multichannel profiles. Our methodology is inspired by the functional Principal Component Analysis (PCA), which has been successfully applied by Paynabar et al. [40], Grasso et al. [18], and Paynabar et al. [41] to deal with intrinsic inner- and inter-channel correlations of profiles. These existing methods follow the standard PCA approach to select a few principal components (projections or eigenvectors) that contain a large amount of variation or information in the profile data under the normal operational or in-control state. This kind of dimension reduction approach might be reasonable from the estimation or curve fitting/smoothing viewpoint under the in-control state, but unfortunately it is ineffective in the context of process monitoring, especially for multivariate or multichannel profiles. This is because it does not reflect the possible change, and often fail to capture the profile information under the out-of-control state. Here we propose to develop a PCA method that

can automatically take into account the change information under the out-of-control state.

Note that there are two different phases of profile monitoring: one is Phase I for offline analysis when a retrospective data set is used to estimate and refine the underlying model and its parameters, and the other is Phase II when the estimated model in Phase I is used for online process monitoring. Here we focus on the Phase I analysis, and hopefully our results can shed new light for Phase II monitoring of multichannel profiles as well. In addition, we should acknowledge that the importance of dimension reduction and feature selection for high-dimensional data via thresholding or shrinkage is well-known in modern statistics, including the profile monitoring literature. Jeong et al. [25] incorporated the hard thresholding into the Hotelling T^2 statistics in the context of online monitoring of single profiles, and Jeong et al. [26] proposed a hard thresholding method to obtain projection information by optimizing "overall relative reconstruction error". Zou et al. [62] applied LASSO shrinkage in linear model coefficients for online monitoring linear profiles problem. However, these existing methods use thresholding or shrinkage to conduct *one-shot* dimension reduction, whereas our proposed methodology splits the dimension reduction process into two steps using two different methods: PCA for the in-control state, and soft-thresholding for the out-of-control state.

The remainder of this paper is organized as follows. In Section 2.2, we present the mathematical formulation of multichannel profile monitoring. In Section 2.3, we propose our thresholded PCA method, and provide a guideline on how to select the corresponding tuning parameters. In Section 2.4, we use the real forging process data and simulations to illustrate the usefulness of our proposed thresholded PCA method. Concluding remarks and future research directions are presented in Section 2.5.

2.2 Problem Formulation and Background

Suppose that a random sample of m multichannel profiles, each with p channels, is collected from a production process. Mathematically, each of the m multichannel profile observations is a p-dimensional curve denoted by $\mathbf{X}_i(t) = (X_i^{(1)}(t), ..., X_i^{(p)}(t))^T$, where $t \in [0, 1]$, for $i = 1, \cdots, m$. We assume that the process is initially in-control and at some unknown time

τ, the process may become out-of-control in the sense of the mean shifts of the profiles $\mathbf{X}_i(t)$'s. Specifically, we assume that the data are from the change-point additive noise model

$$\mathbf{X}_i(t) = \begin{cases} \boldsymbol{\mu}_1(t) + \mathbf{Y}_i(t), & \text{when } i = 1, ..., \tau, \\ \boldsymbol{\mu}_2(t) + \mathbf{Y}_i(t), & \text{when } i = \tau + 1, ..., m, \end{cases} \qquad \text{for } 0 \le t \le 1, \tag{2.2.1}$$

for some unknown $0 \le \tau < m$, where the $\mathbf{Y}_i(t)$'s are independent and identically distributed (i.i.d.) p-dimensional "noise" curves with mean $\mathbf{0}$, i.e., $\mathbf{Y}_i(t) = (\mathbf{Y}_i^{(1)}(t), \cdots, \mathbf{Y}_i^{(p)}(t))^T$ and $\mathbf{E}(\mathbf{Y}_i^{(j)}(t)) = 0$ for all dimension $j = 1, \cdots, p$ and for all observations $i = 1, \cdots, m$.

In Phase I profile monitoring, $\boldsymbol{\mu}_1(t)$ and $\boldsymbol{\mu}_2(t)$ are two unknown p-dimensional mean functions, and we want to utilize the observed $\mathbf{X}_i(t)$'s to test the null hypothesis $H_0 : \boldsymbol{\mu}_1(t) = \boldsymbol{\mu}_2(t)$ (i.e., no change) against the composite alternative hypothesis $H_a : \boldsymbol{\mu}_1(t) \neq \boldsymbol{\mu}_2(t)$ (i.e., a change occurs at some unknown time). In addition, we also impose the classical Type I probability error constraint

$$P_{H_0}(\text{reject } H_0 : \boldsymbol{\mu}_1(t) = \boldsymbol{\mu}_2(t)) \le \alpha, \tag{2.2.2}$$

for some pre-specified constant α, e.g., $\alpha = 5\%$.

To test the hypothesis $H_0 : \boldsymbol{\mu}_1(t) = \boldsymbol{\mu}_2(t)$ under model (2.2.1) subject to the Type I error constraint in (2.2.2), it is important to make suitable assumptions of the correlation of both within and between profile channels. To characterize these correlations, as in [41], we apply Karhunen-Loeve expansion theorem to the p-dimensional noise curves $\mathbf{Y}_i(t)$: there exists a set of orthonormal (orthogonal and unit norm) basis functions $\mathcal{V} = \{v_k(t) \in L_2[0,1], k = 1, 2, ...\}$, such that

$$\mathbf{Y}_i(t) = \sum_{k \in \mathcal{V}} \boldsymbol{\xi}_{ik} v_k(t), \qquad \text{for } i = 1, ..., m, \tag{2.2.3}$$

where the number of elements of $\mathcal{V}$ could be either finite or infinite, and the coefficient $\boldsymbol{\xi}_{ik} = (\xi_{ik1}, \cdots, \xi_{ikp})$ is a p-dimensional vector. The key assumption we made is that the coefficients $\{\boldsymbol{\xi}_{ik}\}$'s are i.i.d. p-dimensional random vectors with mean $\mathbf{0}$ and covariance matrix Σ_k over all $i = 1, \cdots, m$ data points for each base $k \in \mathcal{V}$. Under this assumption, it is evident from (2.2.3) that the $p \times p$ covariance matrix Σ_k satisfies

$$\Sigma_k = \mathbf{E}(\boldsymbol{\xi}_{ik}\boldsymbol{\xi}_{ik}^T) = \mathbf{E}\{\int_0^1 \mathbf{Y}_i(t) v_k(t) dt \int_0^1 \mathbf{Y}_i(t)^T v_k(t) dt\}, \tag{2.2.4}$$

since the basis functions $v_k(t)$'s are orthonormal for each $k \in \mathcal{V}$.

It is useful to briefly discuss the implication of (2.2.3) on the correlations of multichannel profiles. As in the standard functional data analysis, the real-valued basis functions $v_k(t)$'s are closely related to the *inner-channel* correlation of the profiles. Meanwhile, since the p-dimensional curve is decomposed into the same real-valued basis functions $v_k(t)$'s in (2.2.3), the *inter-channel* correlations of the p-channel profiles are characterized by the correlation matrices Σ_k's in (2.2.4) of the coefficients $\{\boldsymbol{\xi}_{ik}\}$'s. In practice, both the basis functions $v_k(t)$'s and the covariance matrices Σ_k's are unknown and needed to be estimated, see next section.

When testing the hypotheses under the change-point model in (2.2.1), it is well-known that the edge effect exists when the true change time τ occurs at the boundary of $[1, m)$, and this can be circumvented by an additional assumption that $\rho_1 m \leq \tau \leq \rho_2 m$ for two constants $0 < \rho_1 < \rho_2 < 1$. To highlight our main ideas, we will not discuss this subtle edge effect here: we will develop our proposed test under the assumption when the change time $1 \leq \tau < m$ is unknown, but will investigate its power properties when both τ and $m - \tau$ are moderately large, e.g., $\tau = [m/2]$.

2.3 Our Proposed Thresholded PCA Methodology

In this section, we propose a thresholded multivariate functional PCA methodology for Phase I monitoring of multichannel profiles. For the purpose of easy understanding, this section is subdivided into three subsections. In Subsection 2.3.1, we review the multivariate functional PCA method that estimates the basis $v_k(t)$'s in (2.2.3) and the covariance matrices Σ_k's in (2.2.4). This allows us to reduce the data from the space of p-dimensional profiles $\mathbf{X}_i(t)$'s to the space of the coefficients $\boldsymbol{\xi}_{ik}$'s in (2.2.3) under the in-control state. In Subsection 2.3.2, our proposed method is developed as a hypothesis test for the change-point model in (2.2.1) augmented by soft-thresholding technique that has a nature semi-Bayesian interpretation and is closely related to the generalized likelihood ratio test. Here the soft-thresholding technique selects significant coefficients $\boldsymbol{\xi}_{ik}$'s in (2.2.3) that are likely affected

by the change, and thus can be thought of as a further dimension reduction under the out-of-control state. In Subsection 2.3.3, based on asymptotic analysis, we provide a guidance on the choice of tuning parameters in our proposed thresholded PCA methodology.

2.3.1 Basis and Covariance Estimation

To have a better understanding of the basis and covariance matrix estimation under the change-point model in (2.2.1), we first consider the estimation under the unrealistic case when the noise functions $\mathbf{Y}_i(t)$'s in (2.2.3) were observable. Recall that the p-dimensional functions $\mathbf{Y}_i(t)$'s are decomposed into the same real-valued basis functions $v_k(t)$'s in (2.2.3), this motivates us to evaluate the *inner-channel* correlation of $\mathbf{Y}_i(t)$'s by the following covariance function:

$$c(t,s) = \mathbf{Cov}\{\mathbf{Y}_i(t), \mathbf{Y}_i(s)\} = \sum_{j=1}^{p} \mathbf{E}(Y_i^{(j)}(t) \cdot Y_i^{(j)}(s)) \qquad \text{for } 0 \le t, s \le 1, \tag{2.3.1}$$

since $\mathbf{Y}_i(t)$ is a p-dimensional function with mean $\mathbf{0}$. When $p = 1$, the covariance function $c(t,s)$ in (2.3.1) is well studied, and it is well-known that the bases $v_k(t)$'s are the eigenfunctions of $c(t,s)$. Below we will show that similar conclusions also hold under our definition of the covariance function $c(t,s)$ in (2.3.1) for the general $p \ge 2$ case.

To see this, since the basis functions $v_k(t)$'s are orthonormal, it follows from (2.2.3) that

$$c(t,s) = \sum_{k=1}^{\infty} \sum_{j=1}^{p} \mathbf{E}[\xi_{ikj}^2] v_k(t) v_k(s),$$

and

$$\int_0^1 c(t,s) v_k(s) ds = \lambda_k v_k(t),$$

where $\lambda_k = \sum_{j=1}^{p} \mathbf{E}[\xi_{ikj}^2]$, and ξ_{ikj} is the j-th component of the p-dimensional random vector $\boldsymbol{\xi}_{ik}$ for $j = 1, \cdots, p$. Hence, the basis $v_k(t)$'s are the eigenfunctions of $c(t,s)$ for any dimension $p \ge 2$.

It suffices to estimate the covariance function $c(t,s)$ in (2.3.1) from the observable profiles $\mathbf{X}_i(t)$. While the noise terms $\mathbf{Y}_i(t)$'s are unobservable, a good news of the change-point additive noise model in (2.2.1) is that the differences $\mathbf{Y}_{i+1}(t) - \mathbf{Y}_i(t) = \mathbf{X}_{i+1}(t) - \mathbf{X}_i(t)$ are

observable for all $1 \leq i \leq m-1$ except $i = \tau$ (the change-point). Thus the covariance function $c(t,s)$ in (2.3.1) can be estimated by $\mathbf{Y}_{i+1}(t) - \mathbf{Y}_i(t)$, which yields the approximation:

$$\widehat{c}(t,s) = \frac{1}{2(m-1)} \sum_{i=1}^{m-1} (\mathbf{X}_{i+1}(t) - \mathbf{X}_i(t))^T (\mathbf{X}_{i+1}(s) - \mathbf{X}_i(s)). \tag{2.3.2}$$

Note that the denominator is $2(m-1)$, and since the $\mathbf{Y}_i(t)$'s are i.i.d. over $i = 1, \cdots, m$, the estimated function $\widehat{c}(t,s)$ in (2.3.2) is consistent under the reasonable regularity assumption of the alternative hypothesis, see Remark #2 in Paynabar et al. [41].

Next, the estimates of basis functions $\widehat{v}_k(t)$'s can be found as the eigenfunctions of $\widehat{c}(t,s)$ in (2.3.2). As for the estimation of the covariance matrix Σ_k in (2.2.4) of coefficients $\boldsymbol{\xi}_{ik}$, we again take advantage of the differences $\mathbf{Y}_{i+1}(t) - \mathbf{Y}_i(t)$ under the change-point additive noise model in (2.2.1), and approximate it by

$$\widehat{\Sigma}_k = \frac{1}{2(m-1)} \sum_{i=1}^{m-1} \int_0^1 \{\mathbf{X}_{i+1}(t) - \mathbf{X}_i(t)\} \widehat{v}_k(t) dt \int_0^1 \{\mathbf{X}_{i+1}(t) - \mathbf{X}_i(t)\}^T \widehat{v}_k(t) dt. \tag{2.3.3}$$

We follow the standard PCA literature to focus on the first d largest eigenvalues of the function $\widehat{c}(t,s)$ in (2.3.2), and consider the corresponding d eigenfunctions $\widehat{v}_k(t)$'s. However, our choice of the actual value of d will be different here. From the dimension reduction viewpoint, the standard PCA methods often reduce the data directly to a low-dimensional space, and thus the value of d is often chosen to be relatively small. Meanwhile, for our proposed method, the dimension reduction process is split into two steps that correspond to the in-control state and the out-of-control state, respectively. The PCA is used only in the first step to reduce the data from the infinitely functional (or super-high-dimensional) space to an intermediate space of R^d, which will be further reduced to a lower-dimensional space in the second step. As a result, the number d of the chosen principal components of the PCA can be moderately large for our proposed method, e.g., fifties or hundreds.

2.3.2 Thresholded PCA for Monitoring

We are ready to present our proposed method that utilizes the observed profiles $\mathbf{X}_i(t)$'s to test $H_0 : \boldsymbol{\mu}_1(t) = \boldsymbol{\mu}_2(t)$ under the change-point additive noise model (2.2.1). Intuitively, it is natural to construct a test statistic based on the estimation of $\boldsymbol{\mu}_1(t) - \boldsymbol{\mu}_2(t)$. This

suggests us to compare the difference of profile sample means before and after a potential change-point $\ell = 1, 2, ..., m-1$,

$$\boldsymbol{\Delta}_\ell(t) = \sqrt{\frac{\ell(m-\ell)}{m}} \left\{ \frac{1}{\ell} \sum_{i=1}^{\ell} \mathbf{X}_i(t) - \frac{1}{m-\ell} \sum_{i=\ell+1}^{m} \mathbf{X}_i(t) \right\}. \tag{2.3.4}$$

Here the term $\sqrt{\ell(m-\ell)/m}$ scales the difference and standardizes the variance of profile difference. Note that the function $\boldsymbol{\Delta}_\ell(t)$ in (2.3.4) would have mean $\mathbf{0}$ when $H_0 : \boldsymbol{\mu}_1(t) = \boldsymbol{\mu}_2(t)$ is true, but have non-zero mean under $H_a : \boldsymbol{\mu}_1(t) \neq \boldsymbol{\mu}_2(t)$ when $\ell = \tau$ (the change-point).

Next, with the estimated orthonormal basis $\widehat{v}_k(t)$'s and estimated covariance matrix $\widehat{\Sigma}_k$ in (2.3.3), we apply the PCA decomposition in (2.2.3) to project the functions $\boldsymbol{\Delta}_\ell(t)$'s in (2.3.4) from the functional space to a d-dimensional space, which essentially conducts a dimension reduction under the in-control state. Specifically, for each candidate change-point $\ell = 1, 2, ..., m-1$, define the projection to each of the first d principal components, $\boldsymbol{\eta}_{\ell k} = \int_0^1 \boldsymbol{\Delta}_\ell(t)\widehat{v}_k(t)dt$, and then compute the corresponding real-valued statistic

$$U_{\ell,k} = \boldsymbol{\eta}_{\ell k}^T \widehat{\Sigma}_k^{-1} \boldsymbol{\eta}_{\ell k} \tag{2.3.5}$$

for $k = 1, 2, \cdots, d$, where $\widehat{\Sigma}_k$ is defined in (2.3.3).

Note that the statistics $U_{\ell,k}$'s in (2.3.5) are motivated from the scenario when the basis $\nu_k(t)$ and Σ_k are known: if the estimates $\widehat{v}_k(t)$ and $\widehat{\Sigma}_k$ are replaced by their true values, it is straightforward from (2.2.1) to show that $\boldsymbol{\eta}_{\ell k} \sim N(\mathbf{0}, \Sigma_k)$ under the null hypothesis $H_0 : \boldsymbol{\mu}_1(t) = \boldsymbol{\mu}_2(t)$ but $\boldsymbol{\eta}_{\ell k} \sim N(\int_0^1 \boldsymbol{\Delta}_\ell(t)v_k(t)dt, \Sigma_k)$ under the alternative hypothesis $H_a : \boldsymbol{\mu}_1(t) \neq \boldsymbol{\mu}_2(t)$. Hence, when the basis $\nu_k(t)$ and Σ_k are known, the $U_{\ell,k}$'s in (2.3.5) are χ_p^2-distributed under H_0 but should be stochastically larger than χ_p^2 under H_a. When the estimates $\widehat{v}_k(t)$ and $\widehat{\Sigma}_k$ are used, we expect that similar conclusions also hold approximately, e.g., whether the value of $U_{\ell,k}$ in (2.3.5) is large or small indicates whether there is a change along the principal component $\widehat{v}_k(t)$ or not.

Finally, our proposed thresholded PCA methodology considers the soft-thresholding transformation of the $U_{\ell,k}$'s in (2.3.5), so as to smooth out those noisy principal component $\widehat{v}_k(t)$'s that do not provide information about the change under the out-of-control state. To

be more rigorous, we propose a test statistic defined by

$$Q_m = \max_{1\le \ell < m} \sum_{k=1}^{d} (U_{\ell,k} - c)^+, \tag{2.3.6}$$

for some pre-specified "soft-thresholding" parameter $c \ge 0$. Here $(u-c)^+ = \max(u-c, 0)$. Then we reject the null hypothesis $H_0 : \boldsymbol{\mu}_1(t) = \boldsymbol{\mu}_2(t)$ if and only if

$$Q_m > L \tag{2.3.7}$$

for some pre-determined threshold L. The choices of the constants c and L will be discussed in more detail in the next section. When $Q_m > L$, we not only claim that there exists a change point, but also can estimate the change point by

$$\widehat{\tau} = \arg \max_{1\le \ell < m} \sum_{k=1}^{d} (U_{\ell,k} - c)^+. \tag{2.3.8}$$

It is informative to provide some high-level insights of the test statistic Q_m in (2.3.6). Since we do not know the true change-point τ, it is natural to maximize (2.3.6) over all candidate change-points $\tau = \ell$ for $1 \le \ell < m$ from the maximum likelihood estimation or generalized likelihood ratio test viewpoints. The summation of the soft-thresholding transformation $(U_{\ell,k} - c)^+$ in (2.3.6) is more fundamental and can be interpreted from the following semi-Bayesian viewpoint. For a given candidate change-point ℓ, let Z_k be the indicator whether the k-th principal component is affected by the change in the out-of-control state or not, for $k = 1, \ldots, d$. Assume that all principal components are independent, and each has a prior probability π getting affected by the changing event. That is, assume that the changing indicators $Z_1, \ldots, Z_d$ are iid with probability mass function $\mathbf{P}(Z_k = 1) = \pi = 1 - \mathbf{P}(Z_k = 0)$. When $Z_k = 1$, the k-th principal component is affected, and $U_{\ell,k}$ in (2.3.5) represents the evidence of possible change in the log-likelihood-ratio scale. Treating Z_k's as the hidden states, and then the joint log-likelihood ratio statistic of Z_k's and $X_{k,n}$ when testing $H_0 : Z_1 = \ldots = Z_d = 0$ (no change) is

$$\begin{aligned} LLR(n) &= \sum_{k=1}^{d} \{Z_k(\log \pi + U_{\ell,k}) + (1 - Z_k)\log(1-\pi)\} - \sum_{k=1}^{d} \log(1-\pi) \\ &= \sum_{k=1}^{d} Z_k \{U_{\ell,k} - \log((1-\pi)/\pi)\}. \end{aligned}$$

Since the Z_k's are unobservable, it is natural to maximize $LLR(n)$ over $Z_1, \ldots, Z_d \in \{0, 1\}$. Hence, the generalized log-likelihood ratio becomes $\sum_{k=1}^{d} \max\{U_{\ell,k} - \log((1-\pi)/\pi), 0\}$, which is exactly our test statistic Q_m in (2.3.6).

We should acknowledge that from the mathematical viewpoint, the multivariate functional PCA-based monitoring method in Paynabar et al. [41] is the special case of Q_m in Paynabar et al. (2.3.6) when the soft-thresholding parameter $c = 0$, which is reasonable in that context because the number d of principal components is small (e.g., $d = 15$). However, our proposed method is a non-trivial extension of Paynabar et al. [41] from the statistical or dimension reduction viewpoint: we consider a moderately large value d of principal components (e.g., $d = 45$), and a suitable choice of the soft-thresholding parameter $c > 0$ in (2.3.6) is essential to conduct another level of dimension reduction to smooth out those principal components that do not provide information of the change under the out-of-control state.

2.3.3 The Choices of Tuning Parameters

There are two tuning parameters in our proposed thresholded PCA methodology based on the test statistic Q_m in (2.3.6): one is the soft-thresholding parameter c in (2.3.6), and the other is the threshold L in (2.3.7). Practically, one needs to determine c first before selecting L, but below we will present the choice of L first for a given c since it is easier to understand from the statistical viewpoint.

In order to find the threshold L for our proposed methodology to satisfy the Type I error probability constraint in (2.2.2), assume, for now, that the constant c in (2.3.6) is given. Then the constraint in (2.2.2) becomes $\mathbf{P}_{H_0}(Q_m > L) \leq \alpha$. Hence, the threshold L should be the upper α quantile of the distribution of Q_m in (2.3.6) for a given c under H_0, and thus it is sufficient to approximate or simulate the distribution of Q_m under H_0.

There are a couple of numerical ways to do so by generating a large number of Monte Carlo simulates of Q_m under H_0. The first one is when there exists "retrospective profiles" dataset that are collected from an in-control process performing under normal operating conditions. Then in each Monte Carlo run, we can randomly select m profiles and compute

the corresponding values of Q_m. Alternatively, when "retrospective profiles" are not available, as suggested in Paynabar et al. [41], one can use the fact that Q_m under H_0 has the same distribution as

$$G_m = \max_{1\le i<m} \sum_{1\le k\le d} [(\frac{(m-i)i}{m})(\bar{\mathbf{z}}_{k,1,i} - \bar{\mathbf{z}}_{k,i+1,m})^T \widehat{\Sigma}_{zk}^{-1}(\bar{\mathbf{z}}_{k,1,i} - \bar{\mathbf{z}}_{k,i+1,m}) - c]^+ \quad (2.3.9)$$

where $\{\mathbf{z}_{k,i}\}$ is a set of independent standard normal multivariate observations of dimension p, $\bar{\mathbf{z}}_{k,\ell_1,\ell_2} = (\ell_2 - \ell_1 + 1)^{-1} \sum_{i=\ell_1}^{\ell_2} \mathbf{z}_{k,i}$, and $\widehat{\Sigma}_{zk} = \frac{1}{2(m-1)} \sum_{i=1}^{m-1} (\mathbf{z}_{k,i+1} - \mathbf{z}_{k,i})(\mathbf{z}_{k,i+1} - \mathbf{z}_{k,i})^T$. Then the threshold L is chosen as the upper α quantile of simulated statistics G_m's.

Let us now discuss the choice of soft-thresholding parameter c in (2.3.6). The baseline choice of c is $c_0 = 0$, which yields the approach of Paynabar et al. [41] for the scenario when the number d of selected principal components is small. Intuitively, when the number d of principal components are large, the soft-thresholding parameter $c > 0$ in (2.3.6) should be large enough to filter out those non-changing bases $\widehat{v}_k(t)$'s, but cannot be too large to remove some changing principal components and lower the signal-to-noise ratios. Hence, a suitable choice of c will depend on the specific H_a and its effects on the basis projections.

Below we will discuss two different heuristic choices of the soft-thresholding parameter $c > 0$. For that purpose, by (2.3.6), we have

$$\mathbf{P}(\sum_{k=1}^{d} (U_{\ell,k} - c)^+ > L) \le \mathbf{P}(Q_m > L) \le \sum_{\ell=1}^{m-1} \mathbf{P}(\sum_{k=1}^{d} (U_{\ell,k} - c)^+ > L), \quad (2.3.10)$$

which becomes $(m-1)\mathbf{P}(\sum_{k=1}^{d} (U_{\ell,k} - c)^+ > L)$, as the data are iid over $\ell = 1, \cdots, m-1$. Hence, from the asymptotic viewpoint, $\mathbf{P}(Q_m > L)$ and $\mathbf{P}(\sum_{k=1}^{d} (U_{\ell,k} - c)^+ > L)$ go to 0 at the same rate when m is fixed. In particular, when the Type I error constraint α goes to 0, the main probability of interest is to estimate

$$\mathbf{P}_{H_0}(\sum_{k=1}^{d} (U_{\ell,k} - c)^+ > L_c), \quad (2.3.11)$$

where L_c is chosen so that this probably $\le \alpha$. Our proposed choices of c correspond to two different methods to approximate the distribution of $\sum_{k=1}^{d} (U_{\ell,k} - c)^+$ under H_0 : one is the central limit theorem (CLT) when c is small, and the other is the extreme theorem when c is large. Since these two methods yield different results on c, we present them separately in Proposition 1, which assumes that χ_p^2 approximation applies to $U_{\ell,k}$'s.

Proposition 1. *Assume that $U_{\ell,k} \sim \chi^2_p$ under H_0, for all $k = 1, ..., d$;*

(a) *(The CLT approximation when c is small). Assume further that under H_a, exactly d_0 out of d principal components are affected in the sense that $U_{\ell,k} \sim \chi^2_p(\delta^2 p) = \boldsymbol{\epsilon}^T_{\ell k}\boldsymbol{\epsilon}_{\ell k}$ with $\boldsymbol{\epsilon}_{\ell k} \sim N(\delta, I_p)$ for $k = 1, ..., d_0$, and $U_{\ell,k} \sim \chi^2_p$ for $k = d_0 + 1, ..., d$. Then when both d_0 and $d - d_0$ are large, an appropriate choice of c is*

$$c_1 = arg\ min_{c\geq 0} \quad \left\{ -\frac{(\mu_c^{(1)} - \mu_c)d_0}{\sqrt{d_0(\sigma_c^{(1)})^2 + (d-d_0)(\sigma_c)^2}} + \frac{\sqrt{d}\sigma_c}{\sqrt{d_0(\sigma_c^{(1)})^2 + (d-d_0)(\sigma_c)^2}} z_\alpha \right\}, \tag{2.3.12}$$

where $\mu_c = \mathbf{E}_0(U_{\ell,k} - c)^+$ and $(\sigma_c)^2 = Var_0(U_{\ell,k} - c)^+$ when $U_{\ell,k} \sim \chi^2_p$; $\mu_c^{(1)} = \mathbf{E}_1(U_{\ell,k} - c)^+$ and $(\sigma_c^{(1)}))^2 = Var_1(U_{\ell,k} - c)^+$ when $U_{\ell,k} \sim \chi^2_p(\delta^2 p)$.

(b) *(The extreme theory approximation when c is large). For fixed p channels, as $d \to \infty$, the soft-thresholding parameter c can be chosen as*

$$c_2 \approx p + 2\log(d). \tag{2.3.13}$$

The detailed proof of this Proposition will be given later in this subsection. Roughly speaking, in Part (a), the c_1 value maximizes the power of the test under the alternative hypothesis H_a subject to the Type I error constraint α in (2.2.2), and the CLT is used to approximate these error probabilities. That is the reason why we need some prior information on d_0 and δ under the alternative hypothesis H_a. In our numerical studies, when such prior information of H_a is not available, our experiences suggest that $\delta = 1$ and $d_0 = d/3$ yield a good robust result under our simulation numerical setting. The rationale of Part (b) is completely different, and is similar to use the following well-known fact to choose the soft-thresholding parameter of $\sqrt{2\log(d)}$ for d iid $N(0,1)$ random variables, see [15],

$$\lim_{d\to\infty} \frac{\max_{1\leq k\leq d} |Z_k|}{\sqrt{2\log(d)}} = 1 \quad \text{almost surely}$$

when the Z_k's are iid $N(0,1)$. Here we extend the critical value from $\sqrt{2\log(d)}$ for the i.i.d. $N(0,1)$-distributed Z_k's to c_2 for the i.i.d. χ^2_p-distributed $U_{\ell,k}$'s for fixed p as $d \to \infty$, and these two critical values are asymptotically equivalent when $p = 1$.

It is also useful to comment on the main assumption of Proposition 1, which is the χ^2_p-approximation for $U_{\ell,k}$'s in (2.3.5). We acknowledge that this holds rigorously only when the basis $\nu_k(t)$'s and Σ_k's are known. However, we shall emphasize that we only need the approximation of the distributions of $U_{\ell,k}$'s in (2.3.5) to derive an approximation choice of the soft-thresholding parameter c. For this reason, the assumption of χ^2_p-approximation for $U_{\ell,k}$'s in (2.3.5) is not bad when the PCA is used to estimate $\nu_k(t)$'s and Σ_k's.

Now let us turn to the proof of Proposition 1.

Proof of Proposition 1: Let us first prove part (a) when the central limit theorem (CLT) is applicable to $\sum_{k=1}^{d}(U_{\ell,k}-c)^+$. This can occur when the soft-thresholding parameter c is small and the number d of bases is large. By the notation in part (a), the terms $(U_{\ell,k}-c)^+$ are i.i.d. with mean μ_c and variance $(\sigma_c)^2$. Hence, $\sum_{k=1}^{d}(U_{\ell,k}-c)^+ \approx N(\mu_c d, \sigma_c^2 d)$ under H_0 for any given ℓ, and the probability in (2.3.11) can be approximated by

$$\mathbf{P}_{H_0}\left(N(0,1) > \frac{L_c - \mu_c d}{\sigma_c\sqrt{d}}\right). \tag{2.3.14}$$

Hence, in order to satisfy the Type I error probability constraint (2.2.2) with small α, the threshold $L = L_c$ can be approximated by $L_c \approx \mu_c d + \sigma_c\sqrt{d}\ z_\alpha$, where $z_u = z$ such that $\mathbf{P}(N(0,1) > z) = u$.

Likewise, we can also derive the relationship of power function of the proposed test. Under the alternative hypothesis H_a with the change time τ, the term $(U_{\tau,k}-c)^+$ has mean μ_c and variance $(\sigma_c)^2$ if the k-th component is unaffected, and has mean $\mu_c^{(1)}$ and variance $(\sigma_c^{(1)})^2$ if affected. Recall that there are d_0 components are affected. When both d_0 and $d-d_0$ are relatively large, the CLT is applicable to both $\sum_{k=1}^{d_0}(U_{\tau,k}-c)^+$ and $\sum_{k=d_0+1}^{d}(U_{\tau,k}-c)^+$. Hence, the power function of the proposed test is of the order of

$$\begin{aligned}
&\mathbf{P}_{H_1}(\sum_{k=1}^{d}(U_{\tau,k}-c)^+ > L_c) = \mathbf{P}_{H_1}(\sum_{k=1}^{d_0}(U_{\tau,k}-c)^+ + \sum_{k=d_0+1}^{d}(U_{\tau,k}-c)^+ > L_c)\\
\approx\ &\mathbf{P}_{H_1}(d_0\mu_c^{(1)} + \sqrt{d_0}\sigma_c^{(1)}Z_1 + \mu_c(d-d_0) + \sqrt{d-d_0}\sigma_c Z_2 > \mu_c d + \sigma_c z_\alpha\sqrt{d})\\
=\ &\mathbf{P}_{H_1}\left(N(0,1) > -\frac{(\mu_c^{(1)}-\mu_c)d_0}{\sqrt{d_0(\sigma_c^{(1)})^2 + (d-d_0)(\sigma_c)^2}} + \frac{\sqrt{d}\sigma_c}{\sqrt{d_0(\sigma_c^{(1)})^2 + (d-d_0)(\sigma_c)^2}} z_\alpha\right),
\end{aligned}$$

where Z_1 and Z_2 are independent $N(0,1)$ random variables, and the last equation is from the fact that $aZ_1 + bZ_2 \sim N(0, a^2+b^2)$. To maximize the power function under H_a, a

natural choice of c is the one that maximizes the above expression, and this leads to the c_1 value in (2.3.12), and thus part (a) holds.

Now let us prove part (b) by using the extreme theory approximation when c is large. In this case, the CLT usually gives a poor approximation to $\sum_{k=1}^{d}(U_{\ell,k}-c)^+$, and we will explore the following facts:

$$\mathbf{P}_{H_0}(\sum_{k=1}^{d}(U_{\ell,k}-c)^+ > L_c) < \mathbf{P}_{H_0}\Big(\max_{1\le k\le d} U_{\ell,k} > c\Big) < \sum_{k=1}^{d}\mathbf{P}_{H_0}\Big(U_{\ell,k} > c\Big) = d\mathbf{P}\big(\chi_p^2 > c\big).$$

Here the first equality follows from the simple fact that $(U_{\ell,k}-c)^+ > 0$ for some $1 \le k \le d$ when $\sum_{k=1}^{d} U_{\ell,k} - c)^+ > L_c > 0$, and the last equality uses the main assumption of the proposition that $U_{\ell,k} \sim \chi_p^2$ under H_0. To satisfy Type I error constraint in (2.2.2), it suffices to find c such that $\mathbf{P}\big(\chi_p^2 > c\big) \approx \frac{\alpha}{d}$ for fixed p and α as $d \to \infty$, i.e., $\log \mathbf{P}\big(\chi_p^2 > c\big) \approx -\log(d)$.

When $p = 1$, for large $c > 0$, we have

$$\mathbf{P}\Big(\chi_1^2 > c\Big) = 2\mathbf{P}(N(0,1) > \sqrt{c}) \approx 2\frac{\phi(\sqrt{c})}{\sqrt{c}} = \frac{2}{\sqrt{2\pi c}}\exp\Big(-\frac{c}{2}\Big),$$

where we use the well-known fact that $\frac{1}{u+1/u}\phi(u) \le \mathbf{P}(N(0,1) > u) \le \frac{1}{u}\phi(u)$ for all $u > 0$. Taking logarithm both sides, we have $c \approx 2\log(d)$ to satisfy Type I error constraint in (2.2.2) for fixed α as d goes to ∞. This is consistent with the well-known fact that $\sqrt{2\log(d)}$ is the critical soft-thresholding value for the d iid $N(0,1)$ random variables, see Fan [15].

Now we need to extend the above arguments from $p = 1$ to any $p > 1$. The crucial step is to approximate $\mathbf{P}(\chi_p^2 > c)$ for large $c > 0$. By Lemma 1 of Inglot and Ledwina [22], we have

$$\frac{1}{2}\mathcal{E}(c) \le \mathbf{P}(\chi_p^2 > c) \le \frac{1}{\sqrt{\pi}}\frac{c}{c-p+2}\mathcal{E}(c),$$

for $p \ge 2$ and $c > p-2$ and

$$\mathcal{E}(c) = \exp\left\{-\frac{1}{2}\Big(c - p - (p-2)\log(c/p) + \log p\Big)\right\}.$$

This implies that $\log \mathbf{P}(\chi_p^2 > c)$ is asymptotically equivalent to $\log \mathcal{E}(c) \approx -(c-p)/2$ for fixed p as $c \to \infty$, also see Theorems 4.1 and 5.1 in Inglot [21]. Thus in order to satisfy Type I error constraint in (2.2.2), we have $c \approx p + 2\log(d)$ if we ignore all non-essential constants when α and p are fixed, and d goes to ∞. Hence, part (b) is proved.

2.4 Case Study

In this section, we apply our proposed thresholded PCA method to the real forging manufacturing process dataset in Figures 5 and 6 in the Introduction. This dataset includes 207 normal profiles under the in-control state and 69 different fault profiles under the out-of-control state. It was analyzed in Paynabar et al. [41] whose method can be thought of as the special case of our proposed method with the specific soft-thresholding parameter $c_0 = 0$. Below the choice of $c_0 = 0$ will be regarded as the baseline method, and we will focus on whether the values of c_1 and c_2 in Proposition 1 for the soft-thresholding parameter c in (2.3.6) can improve the performance or not as compared to the baseline value $c_0 = 0$.

First, we consider a specific case study setting in Paynabar et al. [41] where 207 normal profiles are followed by the 69 fault profiles, i.e., the change-point $\tau = 207$ for the change-point model in (2.2.1), and the baseline method $c_0 = 0$ was shown to successfully detect the true change-point. When our proposed method with either c_1 or c_2 is applied to this setting of real forgoing data, it can also correctly detect the change-point. Our interpretation is that if the change is significantly large, then all reasonable profile monitoring algorithms, including our proposed methods with any of the three c values in (2.3.6), will be able to detect the change correctly.

Below we will conduct extensive simulation studies that focus on detecting different kinds of smaller changes. A high-level description of our simulation setting is as follows. In each Monte Carlo simulation run, we generate $m = 200$ "observable" profiles from the change-point model in (2.2.1). Under the null hypothesis H_0, all $m = 200$ profiles are simulated from the generative model under the in-control state. Under the alternative hypothesis H_a, we will generate $m = 200$ profiles from the change-point model in (2.2.1) with change-point $\tau = 100$, i.e., the first 100 profiles, $\mathbf{X}_1(t), ..., \mathbf{X}_{100}(t)$, are simulated from the generative model under the in-control state, and the last 100 profiles, $\mathbf{X}_{101}(t), ..., \mathbf{X}_{200}(t)$ are simulated from the generative model under one of many different out-of-control states.

Our proposed thresholded PCA methods with three different soft-thresholding parameters c_0, c_1, c_2, are applied to each set of $m = 200$ simulated "observable" profiles. For each method, we choose the threshold L in (2.3.7) to satisfy Type I error constraint $\alpha = 0.05$ by

using the profiles generated from the null hypothesis H_0. For profiles generated under the alternative hypothesis, each method is applied to see whether it is able to correctly detect the change time $\tau = 100$ or not. This process is repeated for 200 times, and the average performances are reported and compared for three different parameters c_0, c_1, c_2. It is important to emphasize that our proposed thresholded PCA methods do not use any information or knowledge of the profile generative models under the in-control or out-of-control states below, which are only used to generate the $m = 200$ "observable" 4-dimensional vector profiles $\mathbf{X}_i(t)$'s.

For better presentation, the remainder of this section is divided into two subsections. In subsection 2.4.1, we use the real profiles and B-splines to present the generative models of profiles under the in-control state and $2 \times 3 \times 7 = 42$ different out-of-control states. This allows us to generate observed profiles $\mathbf{X}_i(t)$'s from the change-point additive noise model in (2.2.1). In subsection 2.4.2, our proposed thresholded PCA methods are applied to the generated profiles $\mathbf{X}_i(t)$'s, and the performance of the values of c_1 and c_2 in Proposition 1 is then compared with that of the baseline value $c_0 = 0$.

2.4.1 Profile Generative Models

The generative model for profiles under the in-control state is built from the 207 normal profiles, $\mathbf{X}_1(t), ..., \mathbf{X}_{207}(t)$, in the real forging dataset, and we propose to do so by using B-splines. To be more specific, we generate an unevenly spaced set of 66 B-spline basis in $[0, 1]$, and after orthogonalization and normalization we obtain basis $B_1(t), ..., B_{66}(t)$ using the "orthogonalsplinebasis" Package in the free statistical software R 3.1.2. Based on our experiences, the choice of 66 basis yields the best tradeoff to balance the fitting of normal profiles and the computational simplicity, but it can easily be changed to another number. Then our proposed generative model for normal profiles is of the form

$$\mathbf{X}(t) = \sum_{i=1}^{66} \widetilde{\boldsymbol{\theta}}_i B_i(t), \tag{2.4.1}$$

where the 4-dimensional vectors $\widetilde{\boldsymbol{\theta}}_i$'s are assumed to be multivariate normally distributed.

It remains to estimate the distribution of $\widetilde{\boldsymbol{\theta}}_i$'s in (2.4.1) under the in-control state from the observed 207 normal profiles in the real forging dataset. Note that in the original

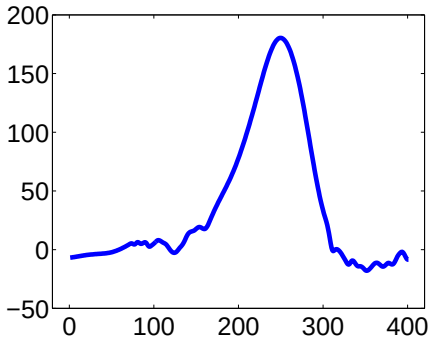

Figure 7: This figure plots the simulated in-control single profile $\mathbf{X}_m^{(1)}(t)$ based on an average of 200 replications. Interval [0, 400] in the x-axis corresponds to $t \in [0/400, 400/400]$.

forgoing dataset, each normal profile $\mathbf{X}_i(t)$ is observed at 1200 different t points, i.e., $t = j/1200$ for $j = 1, \cdots, 1200$. To speed up our computation and to reduce the profile noises, we first apply a non-overlapping moving average function with the window size of 3 to each profile, resulting in 207 "smoothed" 4-dimensional normal profiles $\mathbf{X}(t)$'s with $t \in \left\{\frac{j}{400}; j = 1, ..., 400\right\}$. Next, we fit each of 207 normal profiles $\mathbf{X}(t)$ with B-spline basis $B_1(t), ..., B_{66}(t)$ using the least square estimation method, i.e.,

$$\min_{\boldsymbol{\theta}_1,...,\boldsymbol{\theta}_{66}} \|\mathbf{X}(t) - \sum_{i=1}^{66} \widehat{\boldsymbol{\theta}}_i B_i(t)\|^2.$$

Hence, for each given B-spline basis $i = 1, \cdots, 66$, we obtain 207 fitted values for the 4-dimensional vector $\widehat{\boldsymbol{\theta}}_i$, which allow us to compute the corresponding sample mean and sample covariance matrix, denoted by $\boldsymbol{\theta}_i$ and $\Sigma_{\theta i}$, respectively. Therefore, when we use (2.4.1) to generate normal profiles under the in-control state, we assume that the in-control distribution of $\widetilde{\boldsymbol{\theta}}_i$ is $N(\boldsymbol{\theta}_i, \Sigma_{\theta i})$. To illustrate that this generative model for normal profiles is reasonable, we simulated 200 normal profiles, and computed the average profiles. Figure 7 plots the average profile for the first channel $X^{(1)}(t)$, and clearly it is consistent with the observed normal profiles in Figure 6.

For profiles under the out-of-control (OC) state, we assume that the generative OC model is the same as (2.4.1) but the means of $\widetilde{\boldsymbol{\theta}}_i$'s might be different. We will consider a total of $2 \times 3 \times 7 = 42$ different OC cases, depending on three different factors. First, we consider two different scenarios, depending on how many components/channels of the 4-dimensional random vector $(\widetilde{\theta}_i^{(1)}, \widetilde{\theta}_i^{(2)}, \widetilde{\theta}_i^{(3)}, \widetilde{\theta}_i^{(4)})$ are involved with the change:

(A) All 4 components/channels have new OC mean; and

(B) Only the first 2 components/channels, $\widetilde{\theta}_i^{(1)}$ and $\widetilde{\theta}_i^{(2)}$ have OC mean.

Note that our proposed methods are not designed for Scenario B, and we run simulation to see their performance nevertheless. Second, we consider three cases, depending on which subset of the 66 different $\widetilde{\boldsymbol{\theta}}_i$ in the model (2.4.1) changes their means, or equivalently, which location or subinterval of $[0, 1]$ changes at the original profile scale of $\mathbf{X}_i(t)$ for $0 \leq t \leq 1$:

(I) a local change of $\widetilde{\boldsymbol{\theta}}_i$ for $30 \leq i \leq 37$, i.e., over the interval $\frac{200}{400} \leq t \leq \frac{300}{400}$;

(II) a local change of $\widetilde{\boldsymbol{\theta}}_i$ for $16 \leq i \leq 29$, i.e., over the interval $\frac{99}{400} \leq t \leq \frac{149}{400}$; and

(III) a global change of $\widetilde{\boldsymbol{\theta}}_i$ for all $1 \leq i \leq 66$, i.e., over the interval $0 \leq t \leq 1$.

Third, we consider seven different magnitudes of the change so as to have a better understanding of the detection power as a function of change magnitudes. Note that given the same magnitude of the change, it is the most difficult to detect the local change of Case (I) (where the peak of the profile occurs), and it is the easiest to detect the global change of Case (III). Hence we assign different magnitudes so that the detection powers of these cases are comparable: we assume that the real-valued mean of an affected component $\widetilde{\theta}_i^{(j)}$'s changes from the in-control value θ_i to the out-of-control value $\theta_i + 0.005 + 0.005 * \Delta$, where we set $\Delta = h + 1$ for local change in Case (I), $\Delta = h$ for local change in Case (II), and $\Delta = 0.1 * h$ for global change in case (III), and where seven different values of h will be considered: $h = 1, 2, \cdots, 7$. In summary, there are $2 \times 3 \times 7 = 42$ OC cases depending on the channel, location, and magnitude of the changes, and all numerical values are inspired from the real forging dataset.

2.4.2 Performance Comparison

In this subsection, we report the performance of our proposed thresholded PCA method with three different choices of the soft-thresholding parameter c, and our objective is to see whether the c_1 and c_2 in Proposition 1 will yield a better performance as compared to the baseline $c_0 = 0$ in the sense of detecting those $2 \times 3 \times 7 = 42$ OC cases.

Table 1: The value of d_0 and soft-thresholding parameters c's

	c_0	d_0	c_1	c_2
OC-case (I)	0	15	4.9	11.6
OC-case (II)	0	9	7.0	11.6
OC-case (III)	0	12	4.5	11.6

In order to have a fair comparison, we fix the number of principal components as $d = 45$ for all three choices of soft-thresholding c values, since on average that will explain more than 90% of the profiles variance. In addition, for each method, we choose the threshold L in (2.3.7) to satisfy Type I error constraint $\alpha = 0.05$. Also our proposed methods were developed under the assumption that all 4 components/channels are affected, and the magnitudes of the changes are unknown. Table 1 lists the specific values of c_0, c_1, c_2 used in our study. Note that the value of $c_0 = 0$ and c_2 do not depend on the location of the change, but the value of c_1 depends on the location of the change.

It is useful to explicitly discuss the value d_0 in Table 1. When computing the c_1 value in Proposition 1, we need to know the value of d_0, the number of affected principal components that are relevant to the change among a total of $d = 45$ principal components. Here the value d_0 in Table 1 is chosen by the following data-driven method: We first obtain $U_{\ell,k}^{H_0}\{k = 1, ..., d\}$'s under H_0 using the simulated in-control profiles and record the value A as the top 10% value of $U_{\ell,k}^{H_0}$'s. Then, we compute $U_{\ell,k}^{H_1}\{k = 1, ..., d\}$'s under H_1 using simulated out-of-control profiles, and count how many $U_{\ell,k}^{H_1}$'s are greater than such threshold A. This gives an estimate of d_0 since it indicates the number of altered $U_{\ell,k}$'s if a specific fault occurs. For the purpose of easy computation and comparison, the out-of-control scenario was conducted when all 4 components of affected $\boldsymbol{\theta}_i$ are changed, and the same d_0 and c_1 values were used in the scenario when only 2 out 4 components are changed.

Figure 8 plots the detection power of our proposed methods with three different choices of soft-thresholding c values as functions of change magnitudes when all 4 components/channels of $\boldsymbol{\theta}_i$ are actually changed simultaneously. The top panel deals with the OC-case (I) where a local change affects the rise, peak, and fall segments of the profiles, and all three

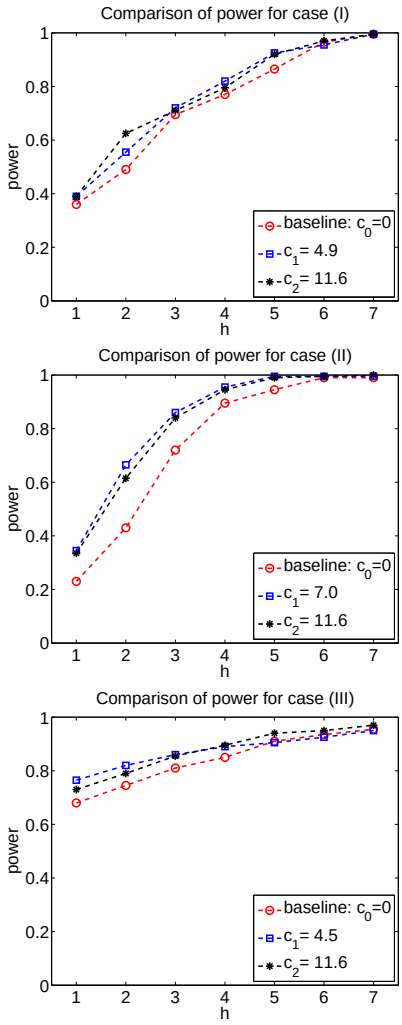

Figure 8: When all 4 channels/components are affected. The three plots correspond to three OC cases, depending on which subset of the 66 different $\widetilde{\boldsymbol{\theta}}_i$ in the model (2.4.1) changes their means. *Upper*: case (I) with a local change for $30 \leq i \leq 37$; *Medium*: case (II) with a local change for $16 \leq i \leq 29$ and *Bottom*: case (III) with a global change for all $1 \leq i \leq 66$. In each figure, each curve represents our proposed method with a specific soft-thresholding c values: Red line with circle (c_0); blue line with square (c_1); and black line with star (c_2). The detection power of each method is plotted as the function of the 7 different change magnitudes.

methods seem to have comparable detection powers, although $c_0 = 0$ is slightly worse. The middle panel shows that under the OC case (II), both c_1 and c_2 can greatly improve the detection power as compared with the baseline $c_0 = 0$, especially when the change magnitude is small (e.g., $h \leq 5$). For large change magnitudes, all three methods have detection power close to 1, implying that all reasonable methods should be able to detect large changes.

A surprising observation of Figure 8 is the bottom panel that considers the OC case (III) when a global change occurs over $[0, 1]$. Intuitively, for a global change, one would expect that the change affects all principal components and hence thresholding might not help. However, the bottom panel of Figure 8 is counter-intuitive, as both c_1 and c_2 seem to yield a larger detection power than $c_0 = 0$, especially for small magnitude h. To gain a deep understanding, Figure 9 plots the box plot of $U_{\ell=100,k}$ under the both IC and OC-case(III) states for all $d = 45$ principal components. From the box plots, for the global change, it is surprising that almost half of $U_{\ell,k}$'s have a similar or smaller median value under OC than IC. We feel that this is the reason why soft-thresholding help improve the detection power in the global change case, as it can filter out those $U_{\ell,k}$'s that have smaller OC values.

We also evaluate the performance of our proposed method in terms of estimating the change-point τ. When the true $\tau = 100$ is estimated as $\widehat{\tau}$, we consider three different measures: $\mathbf{E}(|\widehat{\tau} - \tau|)$, $P(|\tau - \widehat{\tau}| \leq 1)$ (denoted by P1) and $P(|\tau - \widehat{\tau}| \leq 3)$ (denoted by P3). Table 2 reports the Monte Carlo simulation results under these three criteria based on 200 runs. In general all three values c_0, c_1 and c_2 yield comparable results in terms of estimating τ, and it is interesting to note that the thresholding values c_1 and c_2 often have larger P1 and P3 than the baseline $c_0 = 0$ for the OC case (II) with the local-mean shift cases. This suggests that thresholding might be able to locate the small, local change more precisely. One "strange" observation in Table 2 is that $\mathbf{E}(|\widehat{\tau} - \tau|)$ is not necessarily monotone as a function of the change magnitude h. We do not have a deep insight, and one possible explanation is because $|\widehat{\tau} - \tau|$ takes on the integer values, $0, 1, 2, \cdots, 100$, and $\widehat{\tau}$ is a biased estimator.

Figure 10 plots the detection power of our proposed methods when only 2 out of 4 channels/components are affected. It is clear from the top and middle panels of Figure 10

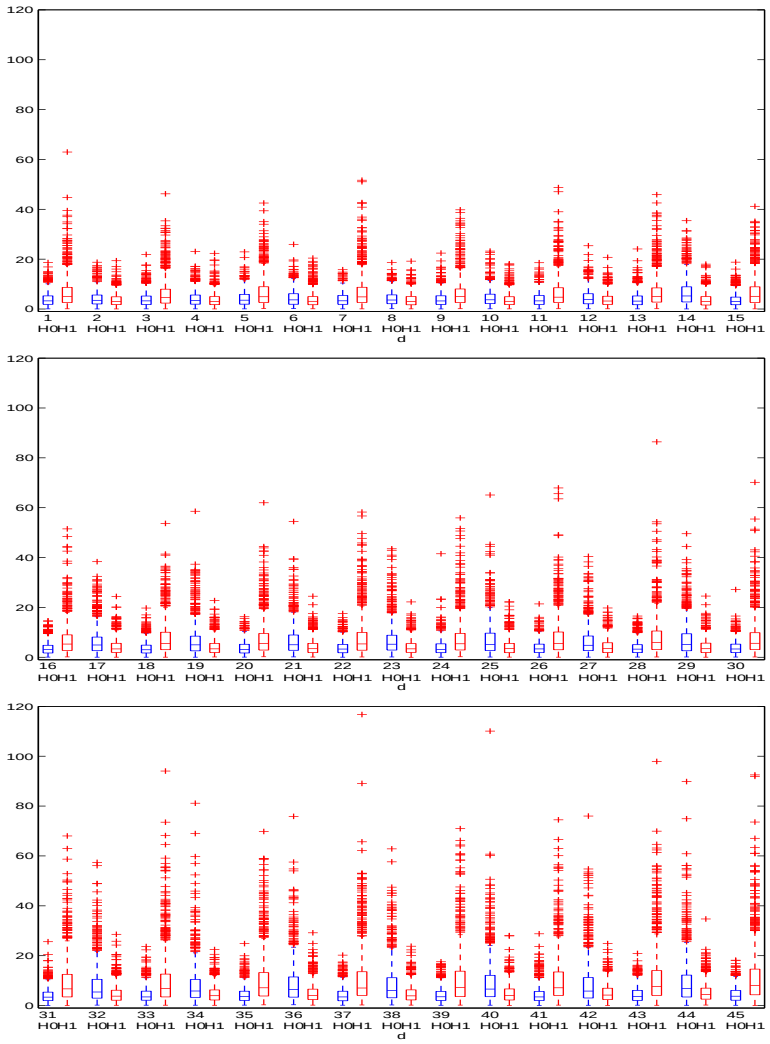

Figure 9: Box plots of $U_{\ell=\tau=100,k}$ under the H_0 hypothesis and H_1 hypothesis for case (III) under all 4 channels affected scenario with $h = 4$ based on 1000 replications. X axis with $k = 1, ..., 45$ represents the projection on the k'th principal components. This plot implies that even for the global change, the OC distribution of the $U_{\ell,k}$'s is not necessarily stochastically larger than those IC distribution over all $k = 1, \cdots, 45$ principal components. We feel that this is the reason why soft-thresholding can improve the detection power in the global change case, as it can filter out those $U_{\ell,k}$'s that have smaller OC values.

Table 2: Comparison of detection biases for each algorithms under 3 different out-of-control cases for all 4 channels affected scenario.

		$\mathbf{E}(\lvert\widehat{\tau}-\tau\rvert)$			$\mathbf{P}(\lvert\tau-\widehat{\tau}\rvert\leq 1)$			$\mathbf{P}(\lvert\tau-\widehat{\tau}\rvert\leq 3)$		
	h	c_0	c_1	c_2	c_0	c_1	c_2	c_0	c_1	c_2
Case (I)	1	5.18 ± 1.71	1.14 ± 1.89	0.86 ± 2.09	0.18	0.15	0.19	0.40	0.44	0.36
	2	1.57 ± 1.33	1.89 ± 1.35	2.65 ± 1.73	0.22	0.22	0.22	0.51	0.50	0.39
	3	0.95 ± 1.21	1.51 ± 1.27	0.59 ± 1.38	0.27	0.26	0.25	0.54	0.54	0.47
	4	0.81 ± 1.10	1.03 ± 1.02	0.59 ± 1.26	0.31	0.36	0.33	0.57	0.63	0.54
	5	0.28 ± 0.86	0.30 ± 0.88	0.09 ± 0.77	0.38	0.36	0.42	0.63	0.63	0.63
	6	0.13 ± 0.73	0.13 ± 0.79	0.58 ± 0.47	0.41	0.42	0.46	0.65	0.68	0.66
	7	0.14 ± 0.54	0.63 ± 0.46	0.29 ± 0.49	0.47	0.46	0.49	0.70	0.72	0.70
Case (II)	1	2.15 ± 2.37	0.16 ± 2.45	2.30 ± 2.30	0.09	0.22	0.22	0.24	0.36	0.36
	2	1.98 ± 1.64	0.78 ± 1.57	0.18 ± 1.50	0.24	0.35	0.35	0.48	0.51	0.53
	3	1.12 ± 0.88	0.76 ± 1.08	1.19 ± 1.23	0.39	0.40	0.42	0.60	0.61	0.63
	4	0.11 ± 0.70	0.67 ± 0.78	0.43 ± 0.71	0.48	0.53	0.56	0.72	0.76	0.76
	5	0.51 ± 0.62	0.02 ± 0.54	0.24 ± 0.57	0.58	0.67	0.65	0.78	0.83	0.86
	6	0.22 ± 0.53	0.51 ± 0.49	0.49 ± 0.49	0.70	0.76	0.73	0.86	0.91	0.90
	7	0.50 ± 0.48	0.02 ± 0.14	0.07 ± 0.16	0.77	0.81	0.80	0.90	0.95	0.95
Case (III)	1	0.07 ± 1.13	0.16 ± 1.07	1.18 ± 1.25	0.35	0.34	0.31	0.57	0.57	0.50
	2	0.58 ± 1.11	0.37 ± 1.01	0.52 ± 1.07	0.39	0.35	0.34	0.60	0.61	0.54
	3	0.85 ± 1.05	0.67 ± 0.94	0.51 ± 0.84	0.43	0.39	0.38	0.64	0.61	0.56
	4	0.15 ± 0.90	0.11 ± 0.73	0.43 ± 0.82	0.45	0.41	0.40	0.67	0.64	0.61
	5	0.13 ± 0.84	0.11 ± 0.66	0.27 ± 0.55	0.47	0.45	0.45	0.69	0.68	0.66
	6	0.44 ± 0.79	0.04 ± 0.58	0.03 ± 0.52	0.49	0.48	0.46	0.70	0.71	0.67
	7	0.39 ± 0.63	0.15 ± 0.54	0.01 ± 0.53	0.51	0.49	0.47	0.72	0.72	0.67

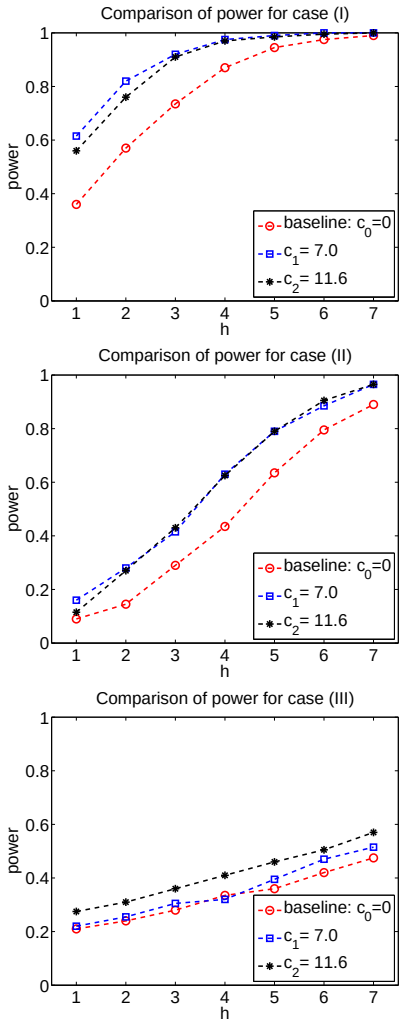

Figure 10: When only 2 out of 4 channels/components are affected. The three plots correspond to three OC cases, depending on which subset of the 66 different $\widetilde{\boldsymbol{\theta}}_i$ in the model (2.4.1) changes their means. *Upper*: case (I) with a local change for $30 \leq i \leq 37$; *Medium*: case (II) with a local change for $16 \leq i \leq 29$ and *Bottom*: case (III) with a global change for all $1 \leq i \leq 66$. In each figure, each curve represents our proposed method with a specific soft-thresholding c values: Red line with circle (c_0); blue line with square (c_1); and black line with star (c_2). The detection power of each method is plotted as the function of the 7 different change magnitudes.

that the c_1 and c_2 values greatly outperforms the baseline $c_0 = 0$ value for almost all shift magnitudes in the OC case of local changes. In the bottom panel for the OC case (III) of the global change, the detection power improvement is significant for c_2 as compared to the baseline $c_0 = 0$. We feel this might be due to the new spatial sparsity where the profile means of only two channels have shifted. While our proposed thresholded PCA method is not designed specifically for the spatial sparsity, the thresholding can actually take care of spatial sparsity to yield better detection power. In addition, as compared to Figure 8, Figure 10 implies that the detection powers when only 2 out of 4 components have changed are less than those when all 4 components have changed.

2.5 Conclusion and Future Work

In this paper, we proposed a thresholded multivariate PCA for multichannel profile monitoring. The novelty of our proposed method is to conduct dimension reduction in two steps: We first apply multivariate PCA to reduce high dimensional multichannel profiles to a reasonable number of features under the in-control state, and then use soft-thresholding techniques to further select informative features under the out-of-control state. These two steps allow us to include only those principal components that are informative to the change and smooth out the noisy ones, thereby yielding efficient monitoring. We also give a couple of suggestions on how to select tuning parameters based on asymptotic analysis. Moreover, we used real forging process dataset and B-splines to build generative methods for multichannel profiles under the in-control state and $2 \times 3 \times 7 = 42$ different out-of-control states, depending on the channel, location, and magnitude of the changes. Our numerical studies demonstrate that the soft-thresholding technique can significantly increase the detection power as compared to the baseline value $c_0 = 0$.

There are a number of interesting problems that have not been addressed here. From the theoretical point of view, it will be useful to investigate the efficiency of our proposed methods, and to find an optimal value of soft-thresholding parameter c that can adaptively adjust for different out-of-control states. Another direction is to investigate how to extend our proposed method to Phase II online profile monitoring. That will be more challenging,

partly because it is more difficult to select informative principal components due to fewer out-of-control profiles since one observes profiles one at a time. Therefore, our research should be interpreted as a starting point for further investigation.

CHAPTER III

GLOBAL OPTIMIZATION OF EXPENSIVE FUNCTIONS USING ADAPTIVE RBF-BASED SURROGATE MODEL VIA UNCERTAINTY QUANTIFICATION

3.1 Introduction

In this paper, we consider the problem of global optimization of expensive functions, i.e., functions which require large computational costs to evaluate. For physical and computational experiments, these functions represent the relationship between input and output variables, and may require days or even weeks to evaluate at a single input setting. One example is the simulation of flow dynamics for rocket engine injectors, which requires numerically solving a large, coupled system of partial differential equations, see Huo and Yang [20]. Even when computation is parallelized over thousands of processing cores, the simulation of a single injector may take months to complete. An important problem about expensive functions is how to optimize the output/response by choosing appropriate settings of the input variables. This problem can be challenging for two reasons. First, it is not feasible to conduct extensive runs of function evaluations to find the optimal input settings, since each function evaluation is expensive. It is thus desirable to identify the optimal input settings with as few runs as possible. The second challenge comes from the complicated nature of the functional relationship. They are usually regarded as "black-boxes", because there is no explicit relationship between the output and input. Although various local optimization methods are available when the derivatives of the functions are known or can be easily obtained, see Boyd and Vanderberghe [6], such methods are not applicable in the present scenario.

In the literature, a widely used practice for global optimization of expensive functions is to sequentially select input settings for function evaluations based on some criterion. The approach consists of two steps. First, it constructs a surrogate model to approximate

the true function based on all the observed function outputs. The advantage of employing surrogate model is that it can provide predictions at any input settings with much cheaper computation. Second, it identifies a new input setting for function evaluation according to some surrogate model based selection criterion. With this approach, it is feasible to approximate the global optimizer of the true function via the surrogate model optimization. For more details along these lines, see Jones et al. [29], Gutmann [19], Regis and Shoemaker [47], etc.

The primary objective of this paper is to propose a novel global optimization framework for optimizing expensive functions. Our approach is motivated by Regis and Shoemaker [47], in which they utilize Radial Basis Functions (RBF) to build a deterministic surrogate model and guide the selection of the next explored point based on the predicted response and some distance criterion. The rationale of using RBFs is that they can capture the nonlinear trend of functions. However, the RBFs they use are pre-determined and lack the flexibility of modeling. Also, it is less efficient to perform function evaluation from their surrogate model, because they use RBFs in a deterministic way without providing prediction uncertainties. Although a distance criterion is used to avoid trap at local optima, it does not incorporate few response information, and thus tends to select new points more "randomly" . To address these issues, we propose to construct surrogate model with RBFs that are chosen adaptively based on the updated outputs, and to select new points based on surrogate models with quantified uncertainties.

There are other approaches for global optimization of expensive functions in the literature. Jones et al. [29] propose a global optimization scheme by constructing a surrogate model with the kriging method. Our approach is different from theirs in that they make strong assumptions on the correlation structure between explored points while ours does not. A detailed review related to the kriging model in global optimization can be found in Jones [28]. Chen et al. [9] propose a global optimization scheme that builds a mean prediction model with linear basis functions selected from a dictionary of functions, and then imposes a Bayesian structure over the mean model to quantify the uncertainty of the

prediction. Our approach is also different from Chen et al. [9]. Instead of using a predetermined discrete function dictionary with a large number of linear functions, we use a moderate number of RBFs that can be adaptively updated based on observed data.

The paper is organized as follows. In Section 3.2, we give a mathematical formulation of the global optimization problem, and provide a review of the RBFs. In Section 3.3, we present the proposed global optimization framework that utilizes adaptive RBF-based Bayesian surrogate model. In Section 3.4, we conduct simulation studies to validate and compare our proposed method with the method by Regis and Shoemaker [47]. Concluding remarks and future research directions are presented in Section 3.5.

3.2 Problem Formulation and Review of RBFs

Suppose $f(\mathbf{x})$ is an expensive function of interest, where $\mathbf{x} = (x^1, ..., x^p)^T \in V$, and V is a convex domain. The objective is to identify an optimal input setting $\mathbf{x}_{opt}$ that maximizes $f(\mathbf{x})$,

$$\mathbf{x}_{opt} = arg \max_{\mathbf{x} \in V} f(\mathbf{x}). \tag{3.2.1}$$

Because it is not practical to evaluate $f(\mathbf{x})$ over V to search the global maximizer due to the huge computational cost, a well-established practice is to sequentially select a few input settings for function evaluation using a two-step strategy. Suppose a set of N function evaluations $\{(\mathbf{x}_i, f(\mathbf{x}_i))\}_{i=1}^N$ are taken. In step 1, a surrogate model is constructed and the resulting model approximation is denoted by $f_N(\mathbf{x})$. Note that the model approximation may not necessarily be an interpolator of the observed points, i.e., $f_N(\mathbf{x}_i) \neq f(\mathbf{x}_i)$. Unlike the true function $f(\mathbf{x})$, the surrogate model is much cheaper to build and evaluate, and thus it is feasible to predict function values over all $\mathbf{x} \in V$. In step 2, the next input setting $\mathbf{x}_{N+1}$ is selected for function evaluation via certain criterion based on the surrogate model from step 1. Steps 1 and 2 iterate until the total computational budget is met. Then the problem of searching for global maximizer $\mathbf{x}_{opt}$ can be transformed as finding that of the model approximation $f_N(\mathbf{x})$, i.e., approximate $\mathbf{x}_{opt}$ by

$$\widehat{\mathbf{x}}_{opt} = arg \max_{\mathbf{x} \in V} f_N(\mathbf{x}). \tag{3.2.2}$$

In the next section, we will follow this two-step strategy and present our proposed optimization framework in detail.

In the remaining part of this section, we give a brief review of the RBFs, which will be used in the proposed framework for the surrogate model construction. In the literature, the RBF is popularly deployed in applied mathematics and neural networks. See Buhmann [7] and Bishop [5]. Several commonly used functions are: (1) Gaussian functions: $r(\mathbf{x}; \boldsymbol{\mu}, s) = \exp\{-s^2||\mathbf{x} - \boldsymbol{\mu}||^2\}$; (2) generalized multi-quadric functions: $r(\mathbf{x}; \boldsymbol{\mu}, s) = (||\mathbf{x} - \boldsymbol{\mu}||^2 + s^2)^\beta$ with $s > 0, 0 < \beta < 1$; (3) generalized inverse multi-quadratic functions: $r(\mathbf{x}; \boldsymbol{\mu}, s) = (||\mathbf{x} - \boldsymbol{\mu}||^2 + s^2)^{-\beta}$ with $s > 0, \beta > 0$; (4) thin plate spline functions: $r(\mathbf{x}; \boldsymbol{\mu}) = ||\mathbf{x} - \boldsymbol{\mu}||^2 \ln(||\mathbf{x} - \boldsymbol{\mu}||)$. In our work, we will focus on the Gaussian RBFs. The Gaussian RBFs have two type of parameters: the center parameter $\boldsymbol{\mu} \in V$ that determines the location of the RBFs, and the scale parameter s that measures the degree of fluctuation of the function. One advantage of using the Gaussian RBFs over other basis functions is that it can capture different trends of response by choosing different centers and scales. For example, a larger s indicates a more concentrated change in the surface, and vice versa.

3.3 General Global Optimization Framework

In this section, we propose a global optimization framework that utilizes adaptive RBF-based surrogate model via uncertainty quantification. In Section 3.3.1, we propose a novel hierarchical normal mixture Bayesian surrogate model with RBFs to approximate the true function, where the model coefficients are sparsely represented to avoid over-fitting, and the parameters of the RBFs are adaptively updated each time a new point is explored. This allows us to predict the function value at any given candidate point. In Section 3.3.2, we propose a model-guided selection criterion that incorporates the expected improvement (EI) of function prediction and its uncertainties. A new point can then be selected to identify either a more promising area of global maximizer or a more uncertain area for further function evaluation. A summary of algorithm and some discussions will be presented in Section 3.3.3.

3.3.1 Normal Mixture Surrogate Model with RBFs

Suppose we observe N explored points $\mathcal{P}_{exp} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, and its function values $y = (y_1, ..., y_N)^T = (f(\mathbf{x}_1), ..., f(\mathbf{x}_N))^T$. Without loss of generality, we assume $E(y_i) = 0$, as otherwise we can approximate $(y_i - \bar{y})$'s instead of y_i's. We propose to construct a surrogate model by a summation of N Gaussian RBFs $r(\mathbf{x}; \boldsymbol{\mu}_i, s_i) = \exp\{-s_i^2||\mathbf{x} - \boldsymbol{\mu}_i||^2\}$ and an error term $\epsilon(\mathbf{x})$.

$$f(\mathbf{x}) = f_N(\mathbf{x}) + \epsilon(\mathbf{x}) = \sum_{i=1}^{N} \beta_i r(\mathbf{x}; \boldsymbol{\mu}_i, s_i) + \epsilon(\mathbf{x}), \tag{3.3.1}$$

Here, an error term is used to model the discrepancy between the model approximation $f_N(\mathbf{x})$ constructed by the RBFs and the true function $f(x)$. We assume that $\epsilon(\mathbf{x})$ follows the normal distribution $\epsilon(\mathbf{x}) \sim N(0, \sigma^2)$. Note that if the center parameters $\boldsymbol{\mu}_i$'s and the scale parameters s_i's are known and fixed, then the surrogate model in (3.3.1) is exactly the same as linear regression.

3.3.1.1 Prior Distributions

Because both $\boldsymbol{\mu}_i$'s and s_i's are unknown, the proposed modeling approach can handle highly nonlinear functions. A uniform prior over a rectangular region is used for $\boldsymbol{\mu} = (\boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_N)$,

$$\boldsymbol{\mu}_i \sim \text{Uniform}(\Omega), i = 1, ..., N, \tag{3.3.2}$$

where $\Omega = \prod_{j=1}^{p} [\min(x_{1:N}^j), \max(x_{1:N}^j))]$, and it is adaptively changed with the addition of new explored points, see [2].

A gamma prior is used for the scale parameters $s = (s_1, ..., s_N)^T$,

$$s_i \sim \text{Gamma}(a_s, b_s), \tag{3.3.3}$$

where a_s and b_s are common to all i's.

We also impose a hierarchical structure on the coefficients β_i's. Define a latent variable $\gamma = (\gamma_1, ..., \gamma_N)^T$ to indicate whether a certain basis function is active or not: $\gamma_i = 1$ indicates that the ith basis is important and should be included in the model, while $\gamma_i = 0$ indicates the opposite. Specifically, we set $\beta_i|(\gamma_i = 0) \sim N(0, \tau_i)$ with small τ_i, and

$\beta_i|(\gamma_i = 1) \sim N(0, C\tau_i)$ with relatively large C. This hierarchical setting is first employed in the Stochastic Search Variable Selection (SSVS) scheme by George and McCulloch [17]. Indeed, it is one type of the "g-prior" (see Zellner [60]) for avoiding over-fitting. Now the mixture normal prior of the model coefficient $\beta = (\beta_1, ..., \beta_N)^T$ can be written as follows:

$$\beta|\gamma \sim N(0, D_r^2), \text{ where } D_r = \text{diag}(a_1\tau_1, ..., a_N\tau_N), \tag{3.3.4}$$

with $a_i = 1$ if $\gamma_i = 0$ and $= C$ if $\gamma_i = 1$, and a binomial prior for the latent variable γ_i,

$$P(\gamma_i = 0) = p_i, P(\gamma_i = 1) = 1 - p_i, i = 1, ..., N. \tag{3.3.5}$$

We also impose an inverse-gamma prior for the residual variance σ^2,

$$\sigma^2 \sim IG(\frac{\nu_0}{2}, \frac{\gamma_0}{2}). \tag{3.3.6}$$

By combining (3.3.1)-(3.3.6), we obtain the full posterior distribution of $\{\beta, \boldsymbol{\mu}, \gamma, \sigma^2, s\}$

$$\begin{aligned} & p(\beta, \boldsymbol{\mu}, \gamma, \sigma^2, s|\mathbf{x}, y) \propto p(y|\beta, \boldsymbol{\mu}, \gamma, \sigma^2, s, \mathbf{x}) \cdot p(\beta|\gamma, \boldsymbol{\mu}) \cdot p(\gamma) \cdot p(s) \cdot p(\boldsymbol{\mu}) \cdot p(\sigma^2) \\ = \quad & \Big[(2\pi\sigma^2)^{-N/2} \exp\{-\frac{1}{2\sigma^2}(y - D(\boldsymbol{\mu}, s) \cdot \beta)^T (y - D(\boldsymbol{\mu}, s) \cdot \beta)\}\Big] \Big[\prod_{i=1}^{N+p} p_i^{\gamma_i}(1 - p_i)^{(1-\gamma_i)}\Big] \\ & \Big[\det(2\pi D_r^2)^{-1/2} \exp\{-\frac{1}{2}\beta^T D_r^{-2}\beta\}\Big] \prod_{i=1}^{N} \Big[\frac{b_s^{a_s}}{\gamma(a_s)} s_i^{a_s - 1} \exp(-b_s s_i)\Big] \Big[\frac{1_\Omega(\boldsymbol{\mu}_{1:N})}{V(\Omega)}\Big] \\ & \Big[(\sigma^2)^{-(\nu_0/2+1)} \exp\{-\frac{\gamma_0}{2\sigma^2}\}\Big], \end{aligned} \tag{3.3.7}$$

where the coefficient matrix $D(\boldsymbol{\mu}, s)$ is defined as

$$D(\boldsymbol{\mu}, s) = \begin{pmatrix} r(\mathbf{x}_1; \boldsymbol{\mu}_1, s_1) & \cdots & r(\mathbf{x}_1; \boldsymbol{\mu}_N, s_N) \\ \vdots & \ddots & \vdots \\ r(\mathbf{x}_N; \boldsymbol{\mu}_1, s_1) & \cdots & r(\mathbf{x}_N; \boldsymbol{\mu}_N, s_N) \end{pmatrix},$$

and the indicator function $1_\Omega(x) = 1$ if $x \in \Omega$, $= 0$ if $x \notin \Omega$.

3.3.1.2 Posterior Sampling

The posterior distribution defined in (3.3.7) is computationally intractable. Markov Chain Monte Carlo (MCMC) method is utilized to solve this problem, see Andrieu et al. [2]

and Koutsourelakis [30]. Specifically, we use the Gibbs sampler to estimate the posterior distribution for the parameters β, γ, σ^2, and the Metropolis-Hasting algorithm to estimate the posterior distribution for the parameters $\boldsymbol{\mu}$ and s, because there is no explicit formula for the posterior distributions of $\boldsymbol{\mu}$ and s. Start with the posterior distributions for β, γ, σ^2. Denote $M = (D(\boldsymbol{\mu}, s)^T D(\boldsymbol{\mu}, s)/\sigma^2 + D_r^{-2})^{-1}$, $h = MD(\boldsymbol{\mu}, s)^T y/\sigma^2$, and $P = I - D(\boldsymbol{\mu}, s)MD(\boldsymbol{\mu}, s)^T/\sigma^2$. Then, the posterior samples of β can be generated by

$$\beta|\boldsymbol{\mu}, \sigma^2, \gamma, s, \mathbf{x}, y \sim \mathrm{N}(h, M). \tag{3.3.8}$$

The posterior samples of σ^2 can be generated by

$$\sigma^2|\beta, \boldsymbol{\mu}, \gamma, s, \mathbf{x}, y \sim \mathrm{IG}(\frac{\nu_0 + N}{2}, \frac{\gamma_0 + |y - D(\boldsymbol{\mu}, s)\beta|^2}{2}). \tag{3.3.9}$$

The posterior samples of γ_i can be generated by

$$P(\gamma_i = 1|\beta, \boldsymbol{\mu}, \sigma, \gamma_{(-i)}, \mathbf{x}, y) = a/(a + b), \tag{3.3.10}$$

where

$$a = f(\beta|\gamma_i = 1, \gamma_{-i}, \boldsymbol{\mu}) f(\gamma_i = 1, \gamma_{-i}) \propto \det(\Sigma^*)^{-1/2} \exp\{-\frac{1}{2}\beta^T(\Sigma^*)^{-1}\beta\}(1 - p_i)$$

with $\Sigma^* = D_r^{i+}$, and D_r^{i+} is D_r with $\gamma_i = 1$,

$$b = f(\beta|\gamma_i = 0, \gamma_{-i}, \boldsymbol{\mu}) f(\gamma_i = 0, \gamma_{-i}) \propto \det(\Sigma^*)^{-1/2} \exp\{-\frac{1}{2}\beta^T(\Sigma^*)^{-1}\beta\} p_i$$

with $\Sigma^* = D_r^{i-}$, and D_r^{i-} is D_r with $\gamma_i = 0$. And the notation $\gamma_{-i} = (\gamma_1, ..., \gamma_{i-1}, \gamma_{i+1}, ..., \gamma_N)^T$ represents the vector of all γ's except γ_i.

Now we turn to the posterior distribution of the parameters $\boldsymbol{\mu}$ and s. The posterior density form of $\boldsymbol{\mu}_i$ is given by

$$p(\boldsymbol{\mu}_i|\boldsymbol{\mu}_{-i}, \beta, s, \sigma, \mathbf{x}, y) \propto \exp\{-\frac{1}{2\sigma^2}(y - D(\boldsymbol{\mu}, s)\beta)^T(y - D(\boldsymbol{\mu}, s)\beta)\} 1_\Omega(\boldsymbol{\mu}_{1:N}), \tag{3.3.11}$$

where $\boldsymbol{\mu}_{-i} = (\boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_{i-1}, \boldsymbol{\mu}_{i+1}, ..., \boldsymbol{\mu}_N)$ denotes the vector of all $\boldsymbol{\mu}$'s except $\boldsymbol{\mu}_i$. We use the Metropolis-Hasting algorithm to generate posterior samples for $\boldsymbol{\mu}_i$. Specifically, at a new step $(k+1)$, we set the proposed density to be a mixture of two densities, and a temporary

sample $\boldsymbol{\mu}_i^*$ can be obtained from the whole domain Ω with uniform probability, or it can be a perturbation of the current iteration $\boldsymbol{\mu}_i^{(k)}$ within its local neighborhood, i.e.,

$$\begin{aligned} & q_1(\boldsymbol{\mu}_i^*) = \text{Uniform}(\Omega), \text{ with probability } \omega, \\ \text{and} \quad & q_2(\boldsymbol{\mu}_i^*) = N(\boldsymbol{\mu}_i^{(k)}, \sigma_\mu^2) \text{ with probability } 1-\omega. \end{aligned} \tag{3.3.12}$$

And we accept this temporary sample $\boldsymbol{\mu}_i^*$ with the acceptance rate

$$A(\boldsymbol{\mu}_i, \boldsymbol{\mu}_i^*) = \min\{1, (\frac{\exp\{-1/(2\sigma^2)|y - D(\boldsymbol{\mu}^*, s)\beta|^2\}}{\exp\{-1/(2\sigma^2)|y - D(\boldsymbol{\mu}, s)\beta|^2\}}) 1_\Omega(\boldsymbol{\mu}_1, .., \boldsymbol{\mu}_i^*, ..., \boldsymbol{\mu}_N)\}$$

where $\boldsymbol{\mu}^* = (\boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_i^*, ..., \boldsymbol{\mu}_N)^T$.

Similarly, we can use the Metropolis-Hasting algorithm to generate samples of s_i. At step $(k+1)$, we choose a temporary s_i^* as a perturbation of the current sample $s_i^{(k)}$ by the proposed density

$$q_3(s_i^*) = N(s_i^{(k)}, \sigma_s^2). \tag{3.3.13}$$

And we accept such sample s_i^* with the acceptance rate

$$A(s_i, s_i^*) = \min\{1, (\frac{\exp\{-1/(2\sigma^2)|y - D(\boldsymbol{\mu}, s^*)\beta|^2\}}{\exp\{-1/(2\sigma^2)|y - D(\boldsymbol{\mu}, s)\beta|^2\}} \cdot \frac{(s_i^*)^{a_s - 1}\exp(-b_s s_i^*)}{s_i^{a_s-1}\exp(-b_s s_i)})\}$$

where $s^* = (s_1, ..., s_i^*, ..., s_N)$.

From (3.3.8)–(3.3.13), we generate posterior samples for $\gamma, \beta, \sigma, \boldsymbol{\mu}, s$ iteratively based the updated estimate for the remaining parameters. For example, we generate $\beta_1^{(k)}$ at step k from $p(\beta_1^{(k)}|\gamma^{(k)}, \beta_{(-1)}^{(k-1)}, \boldsymbol{\mu}^{(k-1)}, \sigma^{(k-1)}, \mathbf{x}, y)$, and generate $\beta_2^{(k)}$ from $p(\beta_2^{(k)}|\gamma^{(k)}, \beta_1^{(k)}, \beta_{3:N}^{(k-1)}, \boldsymbol{\mu}^{(k-1)}, \sigma^{(k-1)}, \mathbf{x}, y)$, etc. Then, the Gibbs sequence

$$\gamma^{(0)}, \beta^{(0)}, \sigma^{(0)}, \boldsymbol{\mu}^{(0)}, s^{(0)}, ..., \gamma^{(k)}, \beta^{(k)}, \sigma^{(k)}, \boldsymbol{\mu}^{(k)}, s^{(k)}, ..., \gamma^{(K)}, \beta^{(K)}, \sigma^{(K)}, \boldsymbol{\mu}^{(K)}, s^{(K)}$$

can be obtained, where K is the total number of iterations. The posterior sample $f_N^{(k)}(\widetilde{\mathbf{x}})$ for model approximation at a candidate explored point $\widetilde{\mathbf{x}}$ can be calculated by

$$f_N^{(k)}(\widetilde{\mathbf{x}}) = \sum_{i=1}^{N} \beta_i^{(k)} r(\widetilde{\mathbf{x}}; \boldsymbol{\mu}_i^{(k)}, s_i^{(k)}).$$

The function prediction $f_N(\widetilde{\mathbf{x}})$ can then be calculated as the average of $f_N^{(k)}(\widetilde{\mathbf{x}})$'s after discarding the first say 40% samples, and the prediction uncertainties can be calculated as the sample variance of $f_N^{(k)}(\widetilde{\mathbf{x}})$'s.

Finally, we note that the mean value of the posterior density of β in (3.3.8) is $h = ((D(\boldsymbol{\mu}, s)^T \cdot D(\boldsymbol{\mu}, s)/\sigma^2 + D_r^{-2})^{-1})D(\boldsymbol{\mu}, s)^T y/\sigma^2$, which is a biased estimator of β with a nugget value D_r^{-2}. Hence, this estimate of β can be regarded as a ridge-type regression estimate. It is deployed to prevent the model coefficients from being too large. Its use can lead to a more stable surrogate model.

3.3.1.3 Tuning Parameters

A remaining issue in the Bayesian computation is the tuning of the hyper-parameters, which is critical for the model performance. For the hyper-parameters related to the RBF, we adopt the settings in Andrieu [2] and Koutsourelakis [30]. Specifically, for the proposed density of the RBF centers $\boldsymbol{\mu}_i$ in (3.3.12), we set $\sigma_\mu^2 = 0.001$. For the prior of the RBF scales s_i in (3.3.3), we set $a_s = 2, b_s = 0$, and for the proposed density of s_i in (3.3.13), we set $\sigma_s^2 = 0.5$. For the hyper-parameters related to model coefficients and residuals, we follow the settings in Chipman et al. [11]. Specifically, for τ_i and C, we suggest to set $\tau_i = \Delta y/(3\Delta \mathbf{x}), C = 50$, where $\Delta x = \max(\mathbf{x}_{1:N}^{1:p}) - \min(\mathbf{x}_{1:N}^{1:p})$, i.e., the largest change in $\mathbf{x}_{1:N}$, and $\Delta y = \sqrt{Var(y)}/5$. For the prior of the indicator variable γ_i, we set $p_i = 0.5$, i.e., the probability of selecting a variable is 50%. For the hyper parameter ν_0 and γ_0 in (3.3.6), we set $\nu_0 = 2$, and $\nu_0\gamma_0$ to be the 99% quantile of the inverse gamma prior that is close to $\sqrt{Var(y)}$.

3.3.2 A New Point Selection Criterion

In this section, we propose to select new explored points that has the best weighted score based on two surrogate-model guided criteria: (1) Expected Improvement (EI) for searching a promising area of global maximizer, and (2) uncertainty of the response prediction for exploring uncertain regions. More precisely, each candidate will be evaluated based on the above two criteria with a score in $[0, 1]$. A desirable point should have a large score in both criteria. Ideally, a good candidate should have both large EI value and large uncertainty.

Because these are two competing criteria, we should consider a weighted version to balance them. A detailed development is given next.

3.3.2.1 EI Criterion

The EI criterion, initially proposed by Mockus and Zilinskas [36], is used to select points close to the global maxima based on a chosen surrogate model. Using this criterion, an explored point is selected to maximize the expected improvement over the best observed response

$$E(I(\mathbf{x})) = E(\max\{y - f_{\max}, 0\}), \tag{3.3.14}$$

where $f_{\max} = \max\{y_1, ..., y_N\}$ is the maximum of the observed model outputs. It is pointed out in Jones et al. [29] that under the Gaussian assumption of $y \sim N(\mu, s_0^2)$, $E(I(\mathbf{x}))$ has the following closed form expression:

$$E(I(\mathbf{x})) = (\mu - f_{\max})\Phi(\frac{\mu - f_{\max}}{s_0}) + s_0\phi(\frac{\mu - f_{\max}}{s_0}). \tag{3.3.15}$$

By examining the terms, we see that the expected improvement is large for those $\mathbf{x}$ having either (i) a predicted value at $\mathbf{x}$ that is much larger than the maximum of outputs obtained so far, i.e., $\mu \gg f_{\max}$, *or* (ii) having much uncertainty about the value of $y(\mathbf{x})$, i.e., when s_0 is large.

In our scenario, since the proposed surrogate model does not satisfy the Gaussian assumption, there is no analytical form for y, and thus it is not practical to calculate $E(I(\mathbf{x}))$ directly. Instead, we calculate the *Sampled Expected Improvement* (SEI) as suggested in Chen et al. [9], i.e., to estimate $E(I(\mathbf{x}))$ based on the posterior samples of y,

$$\widehat{E}(I(\mathbf{x})) = \sum_{k=1}^{K}(\max\{y^{(k)}(\mathbf{x}) - f_{\max}, 0\}), \tag{3.3.16}$$

where $y^{(k)}(\mathbf{x}) = f_N^{(k)}(x)$ is the posterior sample at the kth iteration, and K is the total number of MCMC iterations. Unlike in the Gaussian case, the SEI value in (3.3.16) cannot be expressed as a weighted sum of the improvement term and the prediction uncertainty term. From its definition, only the prediction posterior samples $y^{(k)}(x)$ that are larger than the current best value, f_{max}, are taken in the summation. Thus SEI first identifies

the possible "improvement" area, $\{x|y^{(k)}(x) > f_{max} \text{ for some } k\}$, and then sums over these terms.

Next, we scale the SEI to be within $[0, 1]$. Denote by χ the set of the candidate explored points, which is a discretized grid. Denote $EI_{min} = \min\{\widehat{E}(I(\mathbf{x})), \mathbf{x} \in \chi\}$, $EI_{max} = \max\{\widehat{E}(I(\mathbf{x})), \mathbf{x} \in \chi\}$. Then we can scale $\widehat{E}(I(\mathbf{x}))$ as

$$V_N^{EI}(\mathbf{x}) = \begin{cases} (\widehat{E}(I(\mathbf{x})) - EI_{min})/(EI_{max} - EI_{min}), & \text{if } EI_{max} - EI_{min} \neq 0 \\ 1, & \text{otherwise.} \end{cases}$$

3.3.2.2 Prediction Uncertainties Criterion

Although the SEI criteria can help identify the explored points to expedite the local search for optima, it lacks flexibility in practice in locating the unexplored regions where a global maximizer may exist, and thus may get trapped at local optima. Therefore, we propose to incorporate the prediction uncertainties into the selection criterion for the new explored points. Specifically, we quantify the prediction uncertainties by the 95% confidence interval bandwidth of $f_N(\mathbf{x})$:

$$CIB(f_N(\mathbf{x})) = UCI(f_N(\mathbf{x})) - LCI(f_N(\mathbf{x})), \tag{3.3.17}$$

where $UCI(f_N(\mathbf{x})), LCI(f_N(\mathbf{x}))$ are the upper CI and lower CI calculated as the 97.5% and 2.5% quantiles of the posterior samples $f_N^{(k)}(\mathbf{x})$.

Similarly, denote $CIB_{min} = \min\{CIB(f_N(\mathbf{x})), \mathbf{x} \in \chi\}, CIB_{max} = \max\{CIB(f_N(\mathbf{x})), \mathbf{x} \in \chi\}$. Then $CIB(f_N(\mathbf{x}))$ can be scaled as

$$V_N^{CIB}(\mathbf{x}) = \begin{cases} (CIB(f_N(\mathbf{x})) - CIB_{min})/(CIB_{max} - CIB_{min}), & \text{if } CIB_{max} - CIB_{min} \neq 0 \\ 1, & \text{otherwise.} \end{cases}$$

3.3.2.3 The Weighted Selection Criteria

Now we are ready to present the proposed selection criterion, which is a weighted average of $V_N^{EI}(\mathbf{x})$ and $V_N^{CIB}(\mathbf{x})$,

$$V_N(\mathbf{x}) = (1 - \omega_N)V_N^{EI}(\mathbf{x}) + \omega_N V_N^{CIB}(\mathbf{x}), \tag{3.3.18}$$

where $\omega_N \in (0, 1)$ is a weight coefficient that depends on the number of the current explored points N. A new explored point $\mathbf{x}_{N+1}$ at step $N+1$ is then selected to maximize the selection criterion $V_N(\mathbf{x})$

$$\mathbf{x}_{N+1} = arg \max_{\mathbf{x} \in \chi / P_{exp}} V_N(\mathbf{x}). \tag{3.3.19}$$

Note that a larger ω_N is more likely to lead to new points in a region with large uncertainty, while a smaller ω_N leads to a new point closer to a local optimum. It is usually helpful to explore the whole domain in the beginning when N is small, and refine the approximation in local area when N is large. Thus, the ω value should decrease as N increases. For example, suppose the initial number of explored points is N_{min}, and the total number of explored points is N_{max}. Then a proper weight can be selected as

$$\omega_N = |N - N_{max}|^d / |N_{min} - N_{max}|^d, \tag{3.3.20}$$

where d is a positive constant, say $d = 1$ or 2. When N is small, $\omega_N \approx 1$, the selected $\mathbf{x}_{N+1}$ has larger uncertainty about the value of $y(\mathbf{x})$. When N increases, ω_N decreases, the selected $\mathbf{x}_{N+1}$ has a larger predicted value.

3.3.3 The Proposed Algorithm and Remarks

In the first part of this section, we will present a summary of the algorithm and the flexible usage of the proposed adaptive RBF-based global optimization framework. For abbreviation, we will refer to the proposed method as aRBF. In the second part, we will compare our proposed method with the baseline method proposed in Regis and Shoemaker [47].

Algorithm 1 summarizes our proposed global optimization method by combining the surrogate model construction in Section 3.3.1 and the point selection criterion in Section 3.3.2. Note that the proposed aRBF can be flexibly used in different scenarios. For example, when the number of available function evaluations is small to moderate, there may be not enough observations to estimate all the parameters. In this case, we only need to update some part of the RBF parameters, say the scale parameter s by setting all the scale parameters $s_i \equiv s (i = 1, ..., N)$, and do not update the μ_i parameters. Whether to update all parameters or part of them can be decided based on the magnitude of the

Algorithm 1 Global Optimization Algorithm

1: Choose a small set of initial explored points $P_{exp} = \{\mathbf{x}_1, \mathbf{x}_2, ...\mathbf{x}_{N_{min}}\}$ using a maximin Latin hypercube design
2: **for** $N = N_{\min}, ..., N_{\max}$ **do**
3: Evaluate $f(\mathbf{x})$ on P_{exp}
4: Construct a Bayesian surrogate model $f(\mathbf{x}) = \sum_{i=1}^{N} \beta_i r(\mathbf{x}; \boldsymbol{\mu}_i, s_i) + \epsilon(\mathbf{x})$ as in Section 2.2 based on $\{(\mathbf{x}_i, f(\mathbf{x}_i)), i = 1, ..., N\}$
5: Calculate the SEI in (3.3.16) and uncertainty CIB in (3.3.17), and select a new explored point $\mathbf{x}_{N+1}$ via selection criterion in (3.3.18)-(3.3.19) over $\chi \setminus P_{exp}$
6: Update $P_{exp} = P_{exp} \cup \mathbf{x}_{N+1}$
7: **end for**
8: **Return** the approximated global optimal point $\widehat{\mathbf{x}}_{\text{opt}} = \arg\max_{\mathbf{x}} f_N(\mathbf{x})$

model residuals at the initial stage (i.e., $N \leq N_{\min}$). If updating all parameters leads to relative large model residuals, then we can fix certain parameters instead. The formulas of the posterior distribution in (3.3.7)-(3.3.13) need some minor changes accordingly if certain RBF parameters are fixed. For the above example, one only needs to set s_i in eq. (3.3.7) to be the same s, and update only one s in (3.3.13), and does not need to update the μ_i's in (3.3.11) and (3.3.12).

For the remaining part of this section, we will compare our aRBF with the Global metric stoch-RBF (GRBF) algorithm proposed by Regis and Shoemaker [47] from a theoretical perspective. The GRBF method will be regarded as the baseline method from now on. First we give a brief review. The GRBF employs a surrogate model $s_N(\mathbf{x})$ using RBFs,

$$s_N(\mathbf{x}) = \sum_{i=1}^{N} \lambda_i r(\mathbf{x}; \mathbf{x}_i, s). \tag{3.3.21}$$

The RBFs parameters in (3.3.21) are pre-specified, i.e., the RBF centers are set at the explored points $\mathbf{x}_i$, and s is pre-calculated at the initial stage of optimization. The model coefficients λ_i in (3.3.21) are estimated by solving a deterministic linear system of equation $\Phi\lambda = F$, where $\Phi_{ij} = r(\mathbf{x}_i; \mathbf{x}_j, s)$, $F = (f(\mathbf{x}_1), ..., f(\mathbf{x}_N))^T$. And their point selection

criterion

$$W_N(\mathbf{x}) = (1 - \omega_N^G)V_N^R(\mathbf{x}) + \omega_N^G V_N^D(\mathbf{x}). \tag{3.3.22}$$

is a weighted average of the scaled response prediction $V_N^R(\mathbf{x})$ with

$$V_N^R(\mathbf{x}) = \begin{cases} (s_N(\mathbf{x}) - s_N^{\min})/(s_N^{\max} - s_N^{\min}) \text{ for } s_N^{\max} \neq s_N^{\min}, \\ 1 \text{ o.w,} \end{cases} \tag{3.3.23}$$

and the maximin distance criterion $V_N^D(\mathbf{x})$ with

$$V_N^D(\mathbf{x}) = (d_N(\mathbf{x}) - d_N^{\min})/(d_N^{\max} - d_N^{\min}), \tag{3.3.24}$$

where $s_N^{\max} = \max\{s_N(\mathbf{x})\}$, $s_N^{\min} = \min\{s_N(\mathbf{x})\}$, $d_N(\mathbf{x}) = \min_{1 \leq i \leq N} ||\mathbf{x} - \mathbf{x}_i||^2$, $d_N^{\min} = \min d_N(\mathbf{x})$, $d_N^{\max} = \max d_N(\mathbf{x})$. The ω_N^G can take values in $\{1, 0.8, 0.6, 0.4, 0.2\}$ periodically. For example, if at time $N = 20$, $\omega_N^G = 0.8$, then at the next time $N = 21$, $\omega_N^G = 0.6$. Then a new point $\mathbf{x}_{N+1}$ is selected to maximize $W_N(\mathbf{x})$, and the global maximizer is approximated by $\widehat{\mathbf{x}}_{opt} = \arg\max s_N(x)$.

Although both methods use RBFs, there are two main differences. First, the surrogate model is different. The aRBF uses a Bayesian surrogate model that provides not only predictions but also its uncertainties, while the GRBF utilizes a deterministic surrogate model that only provides predictions. Because our proposed surrogate model is similar to the ridge regression, the approximation of response is more robust and smooth compared to the interpolation surrogate model of the GRBF. The second difference lies in the choice of the selection criterion for new explored points. Both methods utilize a weighted average of two criteria for points selection, one for local refining and the other for global exploration. For the purpose of local refining, we utilize the expected improvement criterion $E(\max\{y - f_{\max}, 0\})$, which can be regarded as a *soft-thresholding* version of $E(y)$. Note that $E(y)$ is the function prediction, which is the same to the $s_N(\mathbf{x})$ in (3.3.21), the surrogate model of GRBF. As previously discussed, thresholding the prediction makes it easier to identify global optimum. For the purpose of global exploration, we select points in (3.3.19) that have larger confidence band (as defined in (3.3.18)), as larger prediction uncertainty is more likely to indicate unexplored regions. In this way, our selection of points utilizes

response information $V_N^{CIB}(\mathbf{x})$ for exploration, and thus is more stable and informative than only using the maximin distance criterion $V_N^D(\mathbf{x})$ as in GRBF. A simulation study will be presented in Section 3.4 to further understand and compare the empirical performance of the two methods.

3.4 Simulation Study

To assess the performance of the aRBF, we compare it with the GRBF, which is regarded as the baseline method. The details of simulation settings are presented in Section 3.4.1 and the simulation results are discussed and summarized in Section 3.4.2.

3.4.1 Simulation Setup

We consider the standard 2d test function "Brainin function", which has been widely used in the global optimization literature, e.g. Jones et al. [29]. The scaled version of "Brainin function" we use here is defined as follows,

$$f(\mathbf{x}) = \frac{-1}{51.95}[(\bar{x}_2 - \frac{5.1\bar{x}_1^2}{4\pi^2} + \frac{5\bar{x}_1}{\pi} - 6)^2 + (10 - \frac{10}{8\pi})\cos(\bar{x}_1) - 44.81], \tag{3.4.1}$$

where $\bar{x}_1 = 15x_1 - 5, \bar{x}_2 = 15x_2$, and $x_1 \in [0,1], x_2 \in [0,1]$. We further restrict this function on an evenly spaced grids $\chi = [0, 0.01, ..., 1]^2$, so that there will be three local maxima and only one global maximum on $[0.96, 0.16]$ with maximum value 1.0473. The contour plot of the Brainin function is given in Figure 11.

The objective is then to find $\mathbf{x}$ that maximizes $f(x)$ in (3.4.1) with as few evaluations as possible. We quantify the efficiency of algorithms by $|\widehat{\mathbf{x}}_{opt} - \mathbf{x}_{opt}|$, the distance between the approximate global maximum $\widehat{\mathbf{x}}_{opt}$ and the true $\mathbf{x}_{opt}$. We randomly choose a small set of $N_{min}(= 16)$ initial explored points using a maximin Latin hypercube design (Santner et al. [51]). Both methods start the same set of $\mathbf{x}_i$'s. Each time the surrogate model is updated by incorporating the f value of a new explored point, we calculate and update the $\widehat{\mathbf{x}}_{opt}$ value. For each algorithm, new explored points are selected and evaluated sequentially until the total number of explored points reaches $N_{max}(= N_{min} + 30) = 46$. This process is repeated 100 times, and the average performances are reported and compared for the two methods.

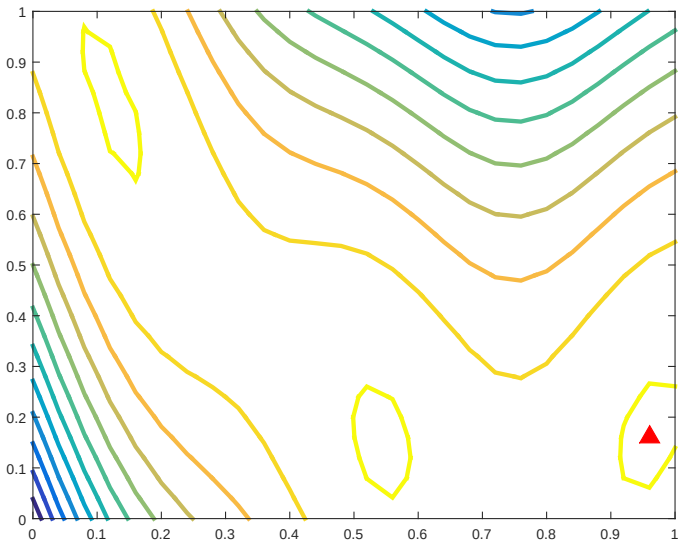

Figure 11: The contour plot of the Brainin function on $[0,1]^2$ with grid size 0.01. The red triangle represents the global optimum.

For fair comparison, we set the initial sampler of RBF parameters in aRBF to be the same as the fixed RBF parameters in GRBF, Specifically, we use Algorithm 1 in Fasshauer and Zhang [16] to select an optimal value of s in GRBF that minimizes a cost function that collects the errors for a sequence of partial fits to the data. The center parameters $\boldsymbol{\mu}_i$'s are set as the explored points $\mathbf{x}_i$'s.

From the simulation study, we find out that for the Brainin test function, updating all parameters in aRBF will lead to relatively large model residuals that do not converge. This might be caused by the small number of function evaluations. Thus we only update one scale parameter s with all $s_i \equiv s$ and fix the center parameter μ_i's to be the explored points. We iterate MCMC 20000 times, and we checked the 20000 iterations of σ^2 to make sure that the MCMC algorithm converges. Also, we discard the first 40% of the samples, and take 1 out of every 5 samples in the remaining 60% of the samples, in order to obtain stable and less correlated posterior samples for model fitting.

3.4.2 Performance Comparison

In this subsection, we first illustrate the proposed aRBF with one particular simulation sample, and then compare it with the GRBF based on 100 simulation samples.

Figure 12 plots the contour of the surrogate model in aRBF and the locations of the

new selected points using aRBF for a simulation sample. It is clearly seen that the first new 10 points are added to explore the unknown region, and after $N \geq 31$, more points are added around the global optimum (i.e., the red triangle).

We report the performance of the aRBF and the GRBF based on 100 random simulations. Our objective is to see whether the aRBF provides a more accurate approximation to the global maximum compared with the GRBF, for the same number of function evaluation. Specifically, we plot in Figure 13 the median value as well as the 10% and 90% quantiles of $|\widehat{\mathbf{x}}_{opt} - \mathbf{x}_{opt}|$ based on the 100 samples for both methods over $N \in [N_{min}, N_{max}]$. Its upper panel is for the GRBF and the lower panel is for the aRBF. In the upper panel, the median value of $|\widehat{\mathbf{x}}_{opt} - \mathbf{x}_{opt}|$ decreases slowly and reaches 0.1 around $N = 43$. The variance of $\widehat{\mathbf{x}}_{opt}$ remains about the same large as N increases, which indicates that the approximation does not converge to the true global maximizer. This is mainly due to the use of the selection criterion in GRBF. Note that the GBRF only uses the distance information among the $\mathbf{x}_i$'s for unknown region exploration without considering the function response. As a result, the selected points spread more randomly, which provides little information on the trend of the function and its peaks. In the lower panel, in the initial stage with $16 \leq N \leq 32$, the median value of $|\widehat{\mathbf{x}}_{opt} - \mathbf{x}_{opt}|$ stays around 0.5, but at $N = 33$, it drops down to 0.1. Also its 10% and 90% quantiles curves are narrower than in the upper panel, suggesting that the variance of $\widehat{\mathbf{x}}_{opt}$ decreases as N becomes large.

Note that our selection criterion is a weighted combination of SEI and CIB with weight function, w_N. From the definition of w_N in (3.3.20), w_N starts from 1 and decreases to 0 as the total number of the explored point, N, becomes large. Hence in the first few iterations of the sequential procedure, it puts more weight on the prediction uncertainty term, CIB, i.e., it focuses on the surrogate fitting. And this may explain why $|\widehat{x}_{opt} - x_{opt}|$ in Figure 13 decreases slowly initially. After certain iterations, w_N gets closer to 0, i.e., $1 - w_N$ is closer to 1. Then it puts more weight on the SEI term in (3.3.18). The purpose of using SEI is to choose the next explored point to potentially improve the search of the optimal point. This may explain why the median value of $|\widehat{x}_{opt} - x_{opt}|$ in Figure 13 drops suddenly after certain iterations. As N becomes large, the weight ω_N in (3.3.20) goes down, and

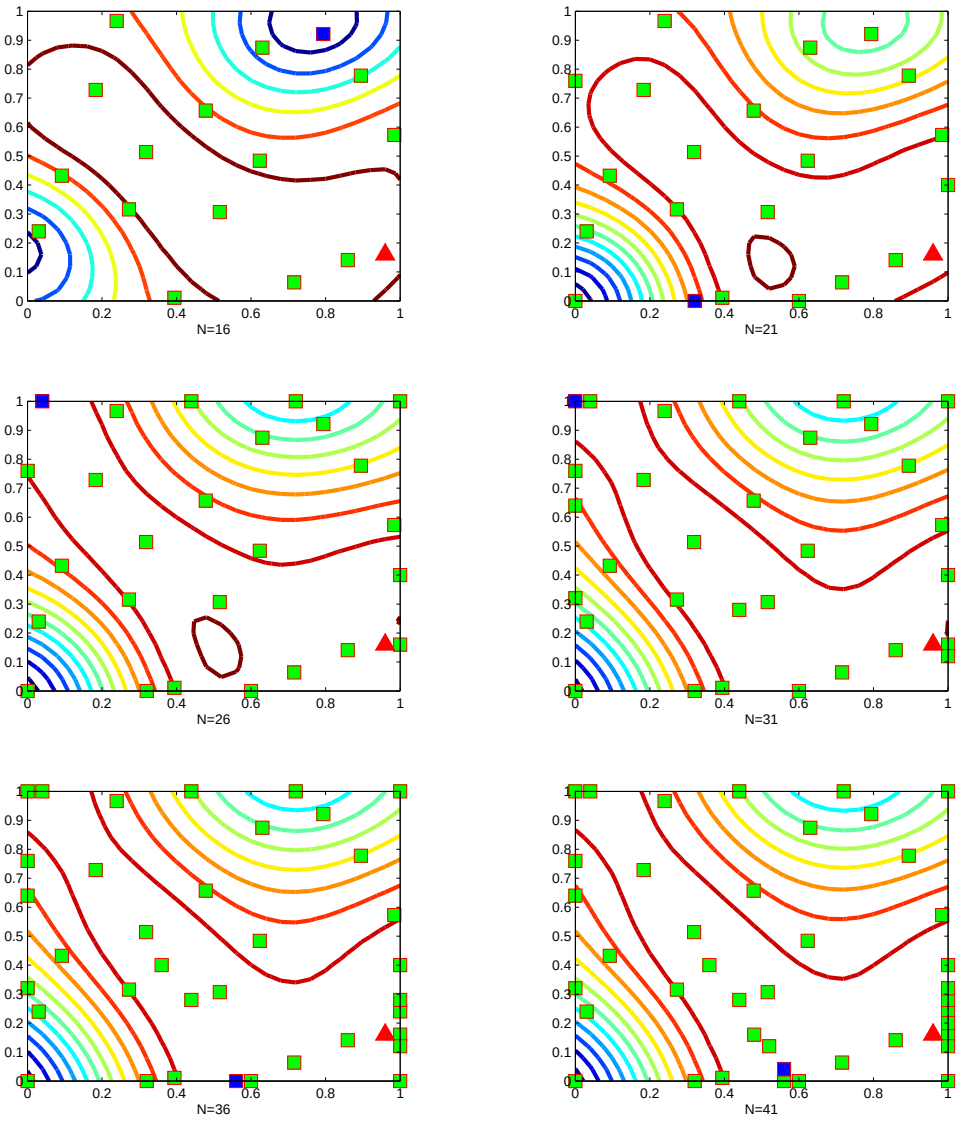

Figure 12: The contour of the surrogate model using the aRBF with the existing explored points (green square), new explored points (blue square), and the global maximum (red triangle) for a simulation sample. Each plot corresponds to a surrogate model with $N = 16, 21, 26, 31, 36, 41$ respectively.

more weight is given to expectation improvement via SEI in (3.3.16), thereby forcing the selected points to focus on refining the local optimum region. Adding the CIB criterion is indeed necessary for exploring more regions and avoid getting trapped in local optima. In other simulations, not presented here, we found that use of EI as the sole selection criterion will keep the values of $|\widehat{\mathbf{x}}_{opt} - \mathbf{x}_{opt}|$ around 0.5 for all $N \in [N_{min}, N_{max}]$, which suggests that the optimization algorithm cannot locate the global maximizer. Therefore, combining both the EI and CIB criteria will be beneficial for searching the global maximum, and can outperform the GRBF.

3.5 Conclusion and Future Work

We have proposed a global optimization framework that iteratively utilizes adaptive RBF-based Bayesian surrogate model to approximate the true function, and to guide the selection of new points for function evaluation. There is novelty in both steps of the optimization strategy. First, the construction of a hierarchical normal mixture surrogate model, where the parameter in the RBFs can be automatically updated to best approximate the true function. Second, the selection criterion for new points by using the EI criterion together with its prediction uncertainty. We have conducted some extensive numerical studies (some not reported here) with standard test functions, and the results demonstrate that the proposed aRBF is more efficient and stable for searching the global maximizer compared with the GRBF.

There are some remaining problems for future research. For example, in the point selection criterion, we predetermine the weight ω_N to balance the trade-off between refining regions and exploring unsampled area. It will be interesting to study how to adaptively select ω_N based on the response values.

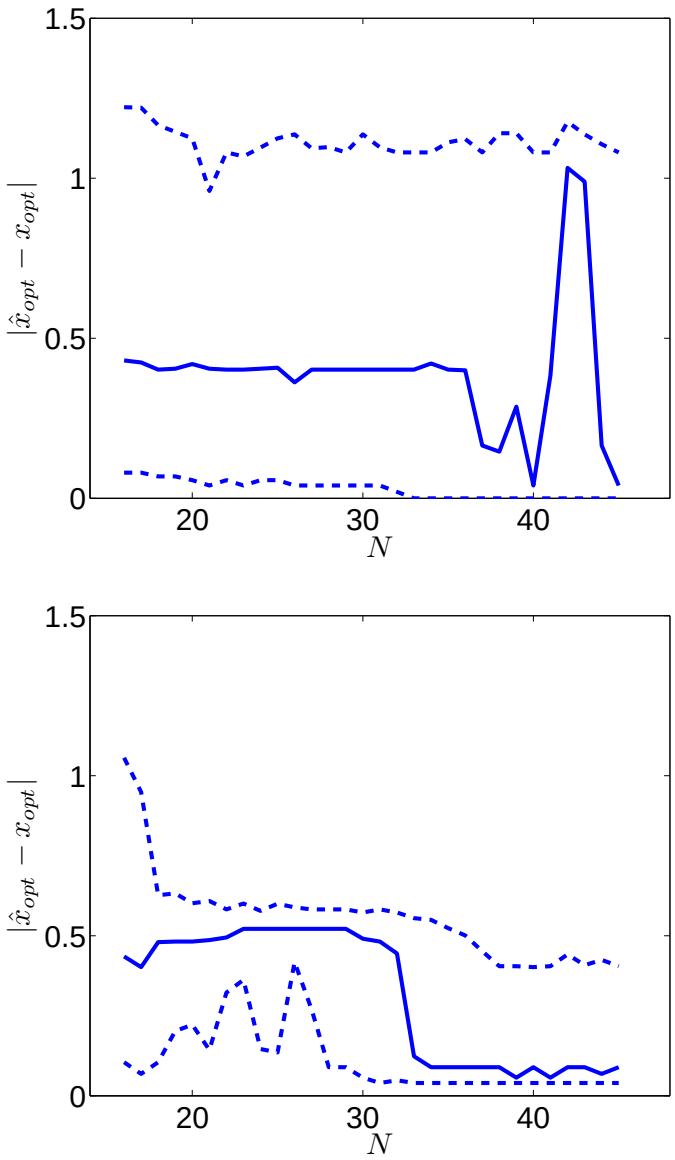

Figure 13: The median value (solid line) as well as 10% and 90% quantiles (dashed line) of $|\widehat{\mathbf{x}}_{opt} - \mathbf{x}_{opt}|$ based on 100 replications. Its upper panel is for the baseline GRBF method and the lower panel is for the proposed aRBF.

REFERENCES

[1] Abdel-Salam, A. S. G., Birch, J. B., and Jensen, W. A., "A semiparametric mixed model approach to phase i profile monitoring," *Quality and Reliability Engineering International*, vol. 29, no. 4, pp. 555–569, 2013.

[2] Andrieu, C., De Freitas, N., and Doucet, A., "Robust full bayesian learning for radial basis networks," *Neural Computation*, vol. 13, no. 10, pp. 2359–2407, 2001.

[3] Basseville, M. and Nikiforov, I. V., *Detection of abrupt changes: theory and application*, vol. 104. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[4] Berkes, I., Gabrys, R., Horváth, L., and Kokoszka, P., "Detecting changes in the mean of functional observations," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 71, no. 5, pp. 927–946, 2009.

[5] Bishop, C. M., *Pattern Recognition and Machine Learning*. New York: NY, Springer, 2006.

[6] Boyd, S. and Vandenberghe, L., *Convex Optimization*. New York: NY, Cambridge University Press, 2004.

[7] Buhmann, M. D., *Radial Basis Functions: Theory and Implementations*, vol. 12. New York: NY, Cambridge University Press, 2003.

[8] Candes, E. J., "Modern statistical estimation via oracle inequalities," *Acta numerica*, vol. 15, pp. 257–325, 2006.

[9] Chen, R.-B., Wang, W., and Wu, C. F. J., "Sequential designs based on bayesian uncertainty quantification in sparse representation surrogate modeling," *Technometrics, to appear*, 2016.

[10] Chicken, E., Pignatiello Jr, J. J., and Simpson, J. R., "Statistical process monitoring of nonlinear profiles using wavelets," *Journal of Quality Technology*, vol. 41, no. 2, pp. 198–212, 2009.

[11] Chipman, H., Hamada, M., and Wu, C. F. J., "A bayesian variable-selection approach for analyzing designed experiments with complex aliasing," *Technometrics*, vol. 39, no. 4, pp. 372–381, 1997.

[12] Cho, H. and Fryzlewicz, P., "Multiple-change-point detection for high dimensional time series via sparsified binary segmentation," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 77, no. 2, pp. 475–507, 2015.

[13] Ding, Y., Zeng, L., and Zhou, S., "Phase i analysis for monitoring nonlinear profiles in manufacturing processes," *Journal of Quality Technology*, vol. 38, no. 3, pp. 199–216, 2006.

[14] Donoho, D. L. and Johnstone, J. M., "Ideal spatial adaptation by wavelet shrinkage," *Biometrika*, vol. 81, no. 3, pp. 425–455, 1994.

[15] Fan, J., "Test of significance based on wavelet thresholding and neyman's truncation," *Journal of the American Statistical Association*, vol. 91, no. 434, pp. 674–688, 1996.

[16] Fasshauer, G. E. and Zhang, J. G., "On choosing optimal shape parameters for rbf approximation," *Numerical Algorithms*, vol. 45, no. 1-4, pp. 345–368, 2007.

[17] George, E. I. and McCulloch, R. E., "Variable selection via gibbs sampling," *Journal of the American Statistical Association*, vol. 88, no. 423, pp. 881–889, 1993.

[18] Grasso, M., Colosimo, B., and Pacella, M., "Profile monitoring via sensor fusion: the use of pca methods for multi-channel data," *International Journal of Production Research*, vol. 52, no. 20, pp. 6110–6135, 2014.

[19] Gutmann, H. M., "A radial basis function method for global optimization," *Journal of Global Optimization*, vol. 19, no. 3, pp. 201–227, 2001.

[20] Huo, H. and Yang, V., "Large eddy simulation of supercritical combustion of liquid oxygen and kerosene of a bi-swirl coaxial injector," *AIAA Paper*, vol. 429, p. 2013, 2013.

[21] Inglot, T., "Inequalities for quantiles of the chi-square distribution," *Probability and Mathematical Statistics*, vol. 30, no. 2, pp. 339–351, 2010.

[22] Inglot, T. and Ledwina, T., "Asymptotic optimality of new adaptive test in regression model," in *Annales de l'IHP Probabilités et statistiques*, vol. 42, pp. 579–590, 2006.

[23] James, W. and Stein, C., "Estimation with quadratic loss," in *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, vol. 1, pp. 361–379, 1961.

[24] Jensen, W. A., Birch, J. B., and Woodall, W. H., "Monitoring correlation within linear profiles using mixed models," *Journal of Quality Technology*, vol. 40, no. 2, pp. 167–183, 2008.

[25] Jeong, M. K., Lu, J. C., and Wang, N., "Wavelet-based spc procedure for complicated functional data," *International Journal of Production Research*, vol. 44, no. 4, pp. 729–744, 2006.

[26] Jeong, M. K., Lu, J. C., Zhou, W., and Ghosh, S. K., "Data-reduction method for spatial data using a structured wavelet model," *International Journal of Production Research*, vol. 45, no. 10, pp. 2295–2311, 2007.

[27] Jin, J. and Shi, J., "Diagnostic feature extraction from stamping tonnage signals based on design of experiments," *Journal of Manufacturing Science and Engineering*, vol. 122, no. 2, pp. 360–369, 2000.

[28] Jones, D. R., "A taxonomy of global optimization methods based on response surfaces," *Journal of global optimization*, vol. 21, no. 4, pp. 345–383, 2001.

[29] Jones, D. R., Schonlau, M., and Welch, W. J., "Efficient global optimization of expensive black-box functions," *Journal of Global optimization*, vol. 13, no. 4, pp. 455–492, 1998.

[30] Koutsourelakis, P. S., "Accurate uncertainty quantification using inaccurate computational models," *SIAM Journal on Scientific Computing*, vol. 31, no. 5, pp. 3274–3300, 2009.

[31] Lai, T. L., "Sequential changepoint detection in quality control and dynamical systems," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 57, no. 4, pp. 613–658, 1995.

[32] Lawson, A. B. and Kleinman, K., "Spatial and syndromic surveillance for public health," 2005.

[33] Lorden, G., "Procedures for reacting to a change in distribution," *The Annals of Mathematical Statistics*, vol. 42, no. 6, pp. 1897–1908, 1971.

[34] Lorden, G. and Pollak, M., "Nonanticipating estimation applied to sequential analysis and changepoint detection," *Annals of statistics*, vol. 33, no. 3, pp. 1422–1454, 2005.

[35] Mei, Y., "Efficient scalable schemes for monitoring a large number of data streams," *Biometrika*, vol. 97, no. 2, pp. 419–433, 2010.

[36] Mockus, J., T. V. and Zilinskas, A., "The application of bayesian methods for seeking the extremum," *Towards Global Optimization*, vol. 2, pp. 117–129, 1978.

[37] Moustakides, G. V., "Optimal stopping times for detecting changes in distributions," *The Annals of Statistics*, vol. 14, no. 4, pp. 1379–1387, 1986.

[38] Noorossana, R., Saghaei, A., and Amiri, A., *Statistical Analysis of Profile Monitoring*, vol. 865. New York: Wiley, 2011.

[39] Page, E., "Continuous inspection schemes," *Biometrika*, vol. 41, no. 1/2, pp. 100–115, 1954.

[40] Paynabar, K., Jin, J., and Pacella, M., "Monitoring and diagnosis of multichannel nonlinear profile variations using uncorrelated multilinear principal component analysis," *IIE Transactions*, vol. 45, no. 11, pp. 1235–1247, 2013.

[41] Paynabar, K., Qiu, P., and Zou, C., "A change point approach for phase-i analysis in multivariate profile monitoring and diagnosis," *Technometrics*, no. forthcoming, pp. 1–37, 2016.

[42] Pollak, M., "Optimal detection of a change in distribution," *The Annals of Statistics*, vol. 13, no. 1, pp. 206–227, 1985.

[43] Pollak, M., "Average run lengths of an optimal method of detecting a change in distribution," *The Annals of Statistics*, vol. 15, pp. 749–779, 1987.

[44] Poor, H. V. and Hadjiliadis, O., *Quickest detection*. New York: Cambridge Univ. Press, 2009.

[45] Qiu, P., *Introduction To Statistical Process Control*. Chapman & Hall/CRC: Boca Raton, FL, 2013.

[46] Qiu, P., Zou, C., and Wang, Z., "Nonparametric profile monitoring by mixed effects modeling," *Technometrics*, vol. 52, no. 3, pp. 265–277, 2010.

[47] Regis, R. G. and Shoemaker, C. A., "A stochastic radial basis function method for the global optimization of expensive functions," *INFORMS Journal on Computing*, vol. 19, no. 4, pp. 497–509, 2007.

[48] Robbins, H. and Siegmund, D., "A class of stopping rules for testing parametric hypotheses," in *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 4, pp. 37–41, The Regents of the University of California, 1972.

[49] Roberts, S., "A comparison of some control chart procedures," *Technometrics*, vol. 8, no. 3, pp. 411–430, 1966.

[50] Rubinstein, R. Y. and Glynn, P. W., "How to deal with the curse of dimensionality of likelihood ratios in monte carlo simulation," *Stochastic Models*, vol. 25, no. 4, pp. 547–568, 2009.

[51] Santner, T. J., Williams, B. J., and Notz, W. I., *The design and analysis of computer experiments*. Springer Science, 2013.

[52] Shiryaev, A. N., "On optimum methods in quickest detection problems," *Theory of Probability & Its Applications*, vol. 8, no. 1, pp. 22–46, 1963.

[53] Siegmund, D., "Change-points: from sequential detection to biology and back," *Sequential Analysis*, vol. 32, no. 1, pp. 2–14, 2013.

[54] Tartakovsky, A. G., Rozovskii, B. L., Blažek, R. B., and Kim, H., "Detection of intrusions in information systems by sequential change-point methods," *Statistical methodology*, vol. 3, no. 3, pp. 252–293, 2006.

[55] Veeravalli, V. V. and Banerjee, T., "Quickest change detection," *Academic press library in signal processing: Array and statistical signal processing*, vol. 3, pp. 209–256, 2013.

[56] Williams, D., *Probability with martingales*. U.K.: Cambridge Univ. press, 1991.

[57] Xie, Y. and Siegmund, D., "Sequential multi-sensor change-point detection," *The Annals of Statistics*, vol. 41, no. 2, pp. 670–692, 2013.

[58] Yakir, B., "A note on the run length to false alarm of a change-point detection policy," *The Annals of Statistics*, pp. 272–281, 1995.

[59] Zamba, K. and Hawkins, D. M., "A multivariate change-point model for statistical process control," *Technometrics*, vol. 48, no. 4, pp. 539–549, 2006.

[60] Zellner, A., "On assessing prior distributions and bayesian regression analysis with g-prior distributions," *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno De Finetti*, vol. 6, pp. 233–243, 1986.

[61] Zhang, N. R., Siegmund, D. O., Ji, H., and Li, J. Z., "Detecting simultaneous changepoints in multiple sequences," *Biometrika*, vol. 97, no. 3, pp. 631–645, 2010.

[62] Zou, C., Ning, X., and Tsung, F., "Lasso-based multivariate linear profile monitoring," *Annals of Operations Research*, vol. 192, no. 1, pp. 3–19, 2012.